Noname manuscript No.
(will be inserted by the editor)

The minimum weighted covariance determinant estimator

Ella Roelant · Stefan Van Aelst · Gert Willems

Received: date / Accepted: date

Abstract In this paper we introduce weighted estimators of the location and dispersion of a multivariate data set with weights based on the ranks of the Mahalanobis distances. We discuss some properties of the estimators like the breakdown point, influence function and asymptotic variance. The outlier detection capacities of different weight functions are compared. A simulation study is given to investigate the finite-sample behavior of the estimators.

Keywords Robust estimation · efficiency · outlier detection

1 Introduction

The classical estimators of the location and scatter of a multivariate data set with n observations and p variables are the sample mean and sample covariance matrix. However, these estimators are not resistent to the presence of outliers in the data set, so robust alternatives that yield reliable estimates even in the presence of contamination are desirable. Since shape or covariance matrices form a cornerstone in multivariate statistical analysis, robust estimators of shape/scatter can be used to construct robust multivariate methods. Such methods have been studied for principal component analysis (Croux and Haesbroeck 2000; Salibian-Barrera et al. 2006), canonical correlation (Croux and Dehon 2002; Taskinen et al. 2006), multivariate regression (Rousseeuw et

The research of Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy)

Ghent University - UGent
Department of Applied Mathematics and Computer Science
Krijgslaan 281 S9
B-9000 Gent
Belgium
Tel.: +32-9-2644756
Fax: +32-9-2644995
E-mail: Ella.Roelant@ugent.be - Stefan.VanAelst@ugent.be - Gert.Willems@ugent.be

al. 2004) and factor analysis (Pison et al. 2003). Pison and Van Aelst (2004) used robust location and scatter estimators to construct diagnostic procedures for multivariate methods.

M-estimators (Maronna 1976) are robust in the sense that they have a bounded influence function but their breakdown point is low. A well-known robust, high-breakdown estimator for location and scatter is the minimum covariance determinant (MCD) estimator, where the estimates are given by the mean and covariance matrix of that half of the data where the smallest determinant of the covariance matrix is attained (Rousseeuw 1984). Butler et al. (1993) studied the asymptotics of the MCD location estimator and Croux and Haesbroeck (1999) discussed the influence function and asymptotic efficiency of the MCD scatter estimator. S-estimators of multivariate location and scatter as studied by Davies (1987) and Lopuhaä (1989) are more efficient, positive breakdown estimators, but their bias can be considerably high. Very efficient, high-breakdown estimators are the classes of constrained M-estimators (Kent and Tyler 1996), τ-estimators (Lopuhäa 1991) and MM-estimators (Tatsuoka and Tyler 2000).

The MCD uses a zero-one weight function. That is, at least half the observations get weight one and the remaining data points get weight zero and thus can be outliers. However, this weight function can make it difficult to identify intermediate outliers which are outliers that are relatively close to the bulk of the data. If such intermediate outliers are not downweighted, then they get the same weight as all other regular points and therefore can attract the estimates such that they become masked. To reveal this masking effect, a more general weight function would be more appropriate. In the regression context, Hössjer (1994) considered robust estimators with weights based on the ranks of the absolute value of the residuals. Moreover, Visek (2001) and Masicek (2004) introduced the least weighted squares estimator as a generalization of the least trimmed squares estimator by using a weight function based on the rank of the squared residuals. In a more general context these estimators can be seen as Weighted Trimmed Likelihood Estimators as discussed by Hadi and Luceño (1997) and Vandev and Neykov (1998). In the same spirit, we consider a generalization of the MCD estimator using weights based on the ranks of the Mahalanobis distances. We specify two types of weight functions (increasing vs non-increasing) which enables us to identify intermediate outliers.

The outline of the paper is as follows. Section 2 defines the estimator. Section 3 describes the algorithm to approximately calculate the estimates. Section 4 discusses the robustness properties of the estimator. We investigate the breakdown point and the influence function at elliptical distributions. Section 5 studies the asymptotic efficiency of the estimates while Section 6 shows results of simulations to investigate the finite-sample behavior of the estimator. Section 7 contains some real data illustrations.

2 The estimator

Let $X_n = \{x_1, \ldots, x_n\}$ be a data set of p-variate observations. We estimate the center μ by minimizing a weighted sum of the squared Mahalanobis distances where the weights depend on the ranks of these distances. We are mainly interested in weight functions $a_n(i) = h^+(i/(n+1))$, $i = 1, \ldots, n$ where $h^+ : (0,1) \to [0, \infty)$ such that

$$\sup\{u; h^+(u) > 0\} = 1 - \alpha,$$

with $0 \leq \alpha \leq \frac{1}{2}$ and $h^+(u) > 0$ for $u \in (0, 1-\alpha]$. Hence, a proportion α of the observations x_i are given weight 0, which ensures that we obtain a robust estimator (see also Hössjer 1994).

Definition 1 The minimum weighted covariance determinant estimator (MWCD) is any solution

$$(\hat{\mu}_{MWCD}(X_n), \hat{V}_{MWCD}(X_n)) = \underset{m,C;\det C=1}{\operatorname{argmin}} D_n(m, C)$$

among all $(m, C) \in \mathbb{R}^p \times \mathrm{PDS}(p)$ where $\mathrm{PDS}(p)$ is the class of positive definite symmetric matrices of size p. The objective function D_n is defined as

$$D_n(m, C) = \frac{1}{n} \sum_{i=1}^{n} a_n(R_i) d_i^2(m, C)$$

with $d_i^2(m, C) = (x_i - m)^T C^{-1} (x_i - m)$ and R_i is the rank of $d_i^2(m, C)$ among $d_1^2(m, C), \ldots, d_n^2(m, C)$.

If there are several solutions to the minimization problem we will arbitrarily choose one as the MWCD estimator. The condition $\det C = 1$ implies that $\hat{V}_{MWCD}$ is an estimator of shape.

In Agulló et al. (2008) it is shown that the MCD can be written as above using the weight function $a_n(i) = I(i \leq k)$ with $n/2 \leq k \leq n$ which explains that the MWCD estimator actually generalizes the MCD estimator by allowing more general weight functions. We expect that the use of different weight functions will give us more insight in the outliers. Both MCD and MWCD have a proportion of the data not contributing to the estimate. In this way, we can accommodate 'strong' outliers. But in contrast to the MCD where each contributing observation has an equal influence, using the MWCD weights allows the influence of these observations to be different. A decreasing weight function puts more weight on points close to the center, while a weight function that is increasing on its support gives higher weight to points further away from the center. Hence, depending on the weight function MWCD treats possible intermediate outliers differently, which enables us to detect them.

An equivalent formulation of the MWCD estimator is obtained as follows.

Proposition 1 *For any data set X_n we have that*

$$\begin{aligned}
&\{(\hat{\mu}(X_n), \hat{V}(X_n)) \in \underset{m,C;\det C=1}{\mathit{argmin}} D_n(m, C)\} \\
&= \{(\tilde{\mu}(X_n), (\det \tilde{\Sigma}(X_n))^{-1/p} \tilde{\Sigma}(X_n)) | (\tilde{\mu}(X_n), \tilde{\Sigma}(X_n)) \in \underset{D_n(m,C)=\tilde{c}}{\mathit{argmin}} \det C \\
&\quad \textit{for any fixed constant } \tilde{c}\}.
\end{aligned}$$

The scatter estimator $\hat{\Sigma}_{MWCD} := \tilde{\Sigma}(X_n)$ can be made a consistent estimator for the covariance matrix of elliptical distributions by selecting the constant $\tilde{c}$ appropriately (see Section 4).

3 Algorithm

We now develop a fast algorithm to calculate an approximate MWCD solution which is similar to the MCD algorithm of Rousseeuw and Van Driessen (1999). We first consider non-increasing weight functions and then propose some modifications for the case of functions that are increasing on their support. The basis of our algorithm is the following proposition which is a generalization of the C-step in Rousseeuw and Van Driessen (1999):

Proposition 2 *Consider a data set* $X_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ *and a non-increasing weight function* a_n. *Denote* $Q_1 := \sum_{i=1}^n a_n(R_{1i}) d_1^2(i)$. *Here* $R_1 = (R_{11}, \ldots, R_{1n})$ *is the rank vector of* $d_1^2(i) = (x_i - \hat{\mu}_1)^T \hat{V}_1^{-1}(x_i - \hat{\mu}_1)$, $i = 1, \ldots, n$ *where* $\hat{\mu}_1 \in \mathbb{R}^p$ *and* $\hat{V}_1 \in \mathbb{R}^{p \times p}$ *with* $\det \hat{V}_1 = 1$. *Now compute the weighted mean and covariance matrix*

$$\hat{\mu}_2 := \hat{\mu}(R_1) = \frac{\sum_{i=1}^n a_n(R_{1i}) x_i}{\sum_{i=1}^n a_n(R_{1i})} \tag{1}$$

$$\hat{\Sigma}_2 := \hat{\Sigma}(R_1) = c_{h^+} \frac{\sum_{i=1}^n a_n(R_{1i})(x_i - \hat{\mu}(R_1))(x_i - \hat{\mu}(R_1))^T}{\sum_{i=1}^n a_n(R_{1i})} \tag{2}$$

where c_{h^+} *is a consistency factor (see Section 4). Denote* $\hat{V}_2 = (\det \hat{\Sigma}_2)^{-1/p} \hat{\Sigma}_2$ *and* $d_2^2(i) = (x_i - \hat{\mu}_2)^T \hat{V}_2^{-1}(x_i - \hat{\mu}_2)$, $i = 1, \ldots, n$ *with corresponding rank vector* R_2. *With* $Q_2 := \sum_{i=1}^n a_n(R_{2i}) d_2^2(i)$ *we then have* $Q_2 \leq Q_1$ *with equality if and only if* $\hat{\mu}_2 = \hat{\mu}_1$ *and* $\hat{V}_2 = \hat{V}_1$.

We plug this generalized C-step in the algorithm of Rousseeuw and Van Driessen (1999) which can be summarized as follows:

1. Start by drawing 1000 random $(p+1)$ subsets J_m of X_n.
2. Compute the corresponding sample mean $\hat{\mu}_m$ and sample shape matrix $\hat{V}_m$. (If $\det(\hat{V}_m) = 0$ for some subset J_m then add points to J_m until $\det(\hat{V}_m) > 0$ or $\#J_m = n$.)
3. For each subset compute the objective function Q_1 based on $(\hat{\mu}_m, \hat{V}_m)$.
4. Apply some C-steps (e.g. two) lowering each time the value of the objective function.
5. Select the 10 subsets which yield the lowest values of the objective function and carry out further C-steps until convergence.
6. The final solution reported by the algorithm is the $\hat{\mu}$ and $\hat{V}$ that correspond to the lowest value of the objective function among these 10.

Note that since there are only a finite number of permutations of the rank vector R, there can only be a finite number of weighted means and covariances as in (1)-(2). Therefore, the uniqueness part of Proposition 2 guarantees that the C-step procedure in Step 5 of the algorithm must converge in a finite number of steps.

If the algorithm finds more than one solution, we arbitrarily choose one of the reported solutions of the algorithm as final solution. Note that there is no guarantee that the algorithm finds all possible solutions for the MWCD estimator.

In the case of an increasing weight function it is not assured anymore that the C-step each time lowers the value of the objective function. Hence, we incorporated in the algorithm that if the C-step does not lower the value of the objective function, then we keep the earlier result as the final solution for that subset (and thus stop applying C-steps). For the 10 best subsets, we set a maximum of 30 C-steps in Step 5 to make sure that the algorithm stops by a given time.

Note that there is no guarantee that the solution reported by the algorithm is an MWCD solution or even a local minimum of the objective function but in our experience the algorithm gives a good approximation of the MWCD solution in most cases.

4 Breakdown point and influence function

We now investigate the robustness properties of the MWCD estimator. The global robustness is investigated by means of the breakdown point while the local robustness is investigated through the influence function.

The breakdown point ε_n^* of an estimator is the smallest fraction of observations from X_n that needs to be replaced by arbitrary values to carry the estimate beyond all bounds (Donoho and Huber 1983). Intuitively, it is clear that for the MWCD this will be approximately α because a proportion α of the observations with largest distances does not affect the estimator. Denote $k = \lfloor (1-\alpha)(n+1) \rfloor$, then k is the number of observations that get a non-zero weight in the MWCD estimator. We assume that the data set X_n satisfies the following condition:

Condition A: No k points of X_n are lying on the same hyperplane of $\mathbb{R}^p$.

Formally, this means that for all $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}$, it holds that $\#\{x_i | \beta^T x_i + \gamma = 0\} < k$ unless β and γ are both zero.

We then have the following proposition.

Proposition 3 *For any data set X_n satisfying condition A it holds that*

$$\varepsilon_n^*(\hat{\mu}_{MWCD}, X_n) = \varepsilon_n^*(\hat{\Sigma}_{MWCD}, X_n) = \frac{\min(n-k+1, k-k(X_n))}{n}$$

with $k(X_n)$ the maximal number of observations of X_n lying on the same hyperplane of $\mathbb{R}^p$.

Since $k = \lfloor (1-\alpha)(n+1) \rfloor$, for data sets in general position (i.e. $k(X_n) = p$), the breakdown point tends to $\min(1-\alpha, \alpha)$.

The MWCD functional $\mathbf{MWCD}$: $\mathcal{H} \to (\mathbb{R}^p \times PDS(p))$ is defined as any solution $\mathbf{MWCD}(H) = (\mu_{MWCD}(H), V_{MWCD}(H))$ to the problem of minimizing

$$D(m, C) = E_H[h^+(G(d_x^2(m, C)))d_x^2(m, C)]$$

subject to

$$\det C = 1$$

with $d_x^2(m, C) = (x-m)^T C^{-1}(x-m)$ and $G(t) = P_H(d_x^2(m, C) < t)$ among all $(m, C) \in \mathbb{R}^p \times PDS(p)$. Note that for general distributions H there is no guarantee that there is a unique solution or even that the number of solutions is finite. If the solution is not unique we arbitrarily select one of the possible solutions. The functional V_{MWCD} corresponds to a shape functional, because its determinant equals 1. It can be easily seen that the resulting MWCD-functional is affine equivariant.

Let us define for any $m \in \mathbb{R}^p$ and $C \in PDS(p)$ the weighted mean and covariance matrix as

$$\mu_{m,C}(H) = \frac{\int h^+(G(d_x^2(m,C)))xdH(x)}{\int h^+(G(d_x^2(m,C)))dH(x)} \tag{3}$$

$$\Sigma_{m,C}(H) = c_{h^+}\frac{\int h^+(G(d_x^2(m,C)))(x-\mu_{m,C})(x-\mu_{m,C})^T dH(x)}{\int h^+(G(d_x^2(m,C)))dH(x)} \tag{4}$$

where c_{h^+} is a constant defined below.

We then have the following result

Proposition 4

$$\{(\mu,V) \in \underset{m,C;\det C=1}{argmin} D(m,C)\}$$
$$\subset \{(\mu,(\det\Sigma)^{-1/p}\Sigma)|(\mu,\Sigma) \in \underset{m=\mu_{m,C};C=\Sigma_{m,C}}{argmin} \det C\}.$$

From Proposition 4 it follows that for every MWCD solution (μ_{MWCD}, V_{MWCD}), the location μ_{MWCD} can be written as a weighted mean as in (3), and for the shape V_{MWCD} there is a corresponding scatter functional Σ_{MWCD} that can be written as a weighted covariance matrix as in (4). The constant c_{h^+} in (4) can be chosen to make the MWCD scatter functional Σ_{MWCD} Fisher-consistent at elliptical model distributions (see Proposition 5).

We now consider estimating the parameters μ and Σ of a model distribution $F_{\mu,\Sigma}$ with density

$$f_{\mu,\Sigma}(x) = \frac{g((x-\mu)^T\Sigma^{-1}(x-\mu))}{\sqrt{\det(\Sigma)}}$$

with $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathrm{PDS}(p)$. The function g is assumed to be known and to have a strictly negative derivative g' so $F_{\mu,\Sigma}$ is an elliptically symmetric, unimodal distribution. At this model distribution, the MWCD scatter functional Σ_{MWCD} becomes Fisher-consistent by setting $c_{h^+} = c_1/c_3$ where

$$c_1 = \frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty h^+(\tilde{G}(r^2))g(r^2)r^{p-1}dr \text{ and } c_3 = \frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty \frac{1}{p}h^+(\tilde{G}(r^2))g(r^2)r^{p+1}dr$$

with $\tilde{G}(t) = P_{F_{0,I}}(X^TX \le t)$. This follows immediately by substituting $F_{0,I}$ for H in (4). We now obtain the following consistency result.

Proposition 5 *The functionals μ_{MWCD} and Σ_{MWCD} are Fisher-consistent for the parameters μ and Σ at elliptical model distributions:*

$$\mu_{MWCD}(F_{\mu,\Sigma}) = \mu \text{ and } \Sigma_{MWCD}(F_{\mu,\Sigma}) = \Sigma.$$

Note that Proposition 5 implies that the resulting functionals μ_{MWCD} and Σ_{MWCD} are unique at elliptical model distributions.

The influence function of a functional T at the distribution H measures the effect on T of an infinitesimal contamination at a single point x (Hampel et al. 1986). If

we denote the point mass at x by Δ_x and consider the contaminated distribution $H_{\varepsilon,x} = (1-\varepsilon)H + \varepsilon\Delta_x$ then the influence function is given by

$$IF(x;T,H) = \lim_{\varepsilon\downarrow 0} \frac{T(H_{\varepsilon,x}) - T(H)}{\varepsilon} = \frac{\partial}{\partial\varepsilon} T(H_{\varepsilon,x})_{|\varepsilon=0}.$$

We will consider the influence function at an elliptical distribution $F_{\mu,\Sigma}$. Due to affine equivariance of $T(H)$ it suffices to look at spherical distributions $F_{0,I}$ with density $f_{0,I}(x) = g(x^T x)$.

Proposition 6 *Denote* $q_\alpha = \tilde{G}^{-1}(1-\alpha)$ *and* $w = h^+ \circ \tilde{G}$, *then*

$$IF(x;\mu_{MWCD}, F_{0,I}) = \frac{w(\|x\|^2)x}{-2c_2} I(\|x\|^2 \leq q_\alpha)$$

with

$$c_2 = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+1} w(r^2) g'(r^2) dr.$$

The influence function of the scatter matrix part Σ_{MWCD} *(for* $p > 1$*) is given by*

$$IF(x;\Sigma_{MWCD}, F_{0,I}) = -\frac{1}{2c_4} xx^T w(\|x\|^2) I(\|x\|^2 \leq q_\alpha) + R(\|x\|) I_p$$

where

$$c_4 = \frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+3} w(r^2) g'(r^2) dr.$$

The term for $R(\|x\|)$ is rather elaborate and can be found in the Appendix (29).

The gross error sensitivity of a functional T at a distribution H is defined as

$$GES(T,H) := \sup_x \|IF(x;T,H)\|.$$

The gross error sensitivity is a measure of the maximal bias caused by an infinitesimal contamination and hence is preferred to be low. In Table 1 we computed the gross-error sensitivity for μ_{MWCD} at the normal model. Throughout the paper we will use the following weight functions: a weight function that is decreasing on its support (MWCD$\downarrow$ estimator), a weight function that is increasing on its support (MWCD$\uparrow$ estimator) and the zero-one weight function which corresponds to the MCD estimator. These functions become zero when $u > 1-\alpha$. For $u \leq 1-\alpha$ we have $h^+_{\text{MWCD}\downarrow}(u) = F^{-1}_{\chi^2_p}\left(1-\frac{u}{2}\right)$ and $h^+_{\text{MWCD}\uparrow}(u) = F^{-1}_{\chi^2_p}\left(\frac{1+u}{2}\right)$. We use the notation MWCD$\downarrow$50 for $\alpha = 0.50$ and MWCD$\downarrow$25 for $\alpha = 0.25$. From Table 1 we see that the MWCD$\downarrow$ estimators have the lowest gross error sensitivities.

Table 1 Gross-error sensitivity at the normal model for the location estimators μ_{MWCD} and μ_{MCD} for different dimensions p and 50% and 25% breakdown point

p	1	2	3	5	10	30
MWCD↓50	6.96	5.75	5.80	6.42	8.18	13.51
MWCD↑50	11.95	9.44	9.11	9.36	10.60	14.73
MCD50	9.46	7.67	7.56	7.98	9.36	13.61
MWCD↓25	2.74	2.86	3.14	3.73	5.02	8.65
MWCD↑25	5.94	5.71	5.87	6.30	7.31	10.20
MCD25	4.16	4.13	4.35	4.85	5.93	8.89

5 Efficiency

If an estimator is Fréchet-differentiable, then its asymptotic variance at the model distribution $F_{0,I}$ can be calculated through its influence function. Neither Fréchet-differentiability nor asymptotic normality have been formally proven for the MWCD estimators. However, we conjecture that the MWCD is Fréchet-differentiable and use this assumption to calculate asymptotic variances through the influence function. For the location estimator μ_{MWCD} we then obtain that the asymptotic variance-covariance matrix equals

$$\mathrm{ASV}(\mu_{MWCD}, F_{0,I}) = E_{F_{0,I}}[IF(x; \mu_{MWCD}, F_{0,I}) \times IF(x; \mu_{MWCD}, F_{0,I})^T]$$

(see e.g. Hampel et al. 1986) which yields

$$\mathrm{ASV}(\mu_{MWCD}, F_{0,I}) = \frac{\int_{\|x\|^2 \le q_\alpha} w(\|x\|^2)^2 \|x\|^2 dF_{0,I}(x)}{4c_2^2}.$$

Similarly, we can calculate the asymptotic variances of the diagonal and off-diagonal elements of the shape matrix V_{MWCD}:

$$\mathrm{ASV}(V_{ii}, F_{0,I}) = \left(2 - \frac{2}{p}\right)\sigma_1 \quad \text{and} \quad \mathrm{ASV}(V_{ij}, F_{0,I}) = \sigma_1$$

with

$$\sigma_1 = \frac{1}{p(p+2)} E_{F_{0,I}}\left[\frac{1}{4c_4^2} w(\|x\|^2)^2 \|x\|^4\right].$$

The asymptotic variances of the diagonal and off-diagonal elements of Σ_{MWCD} become:

$$\mathrm{ASV}(\Sigma_{ii}, F_{0,I}) = 2\sigma_1 + \sigma_2 \quad \text{and} \quad \mathrm{ASV}(\Sigma_{ij}, F_{0,I}) = \sigma_1$$

with

$$\sigma_2 = -\frac{2}{p}\sigma_1 + E_{F_{0,I}}[\gamma^2(\|x\|)] \quad \text{and} \quad \gamma(\|x\|) = \frac{-1}{2c_4} w(\|x\|^2) \frac{\|x\|^2}{p} + R(\|x\|).$$

To gain more insight in the MWCD estimators and how the weighting concept affects their performance, we compare their efficiencies with that of the MCD estimator at the multivariate standard normal distribution $N_p(0, I)$ and multivariate spherical t-distributions t_ν where ν is the degrees of freedom.

For $\alpha = 0.25$, Figure 1a shows the asymptotic relative efficiency (ARE) of the MWCD location estimators, relative to the MCD given by $\text{ARE}(\mu_{MWCD}, \mu_{MCD}) = \text{ASV}(\mu_{MCD})/\text{ASV}(\mu_{MWCD})$ at the multivariate normal distribution. Note that the ARE increases with the dimension p. Moreover, the efficiency of the MWCD↑25 location estimator is comparable to the MCD25 location estimator. Figures 1b and 1c show the ARE at t-distributions with 3 and 8 degrees of freedom, respectively. Figure 1b shows that MWCD↓25 now is more efficient than the MCD25. At the t_8-distribution the MWCD↓25 has the highest ARE and from $p = 5$ on it outperforms the MCD25. Figure 1d shows the ARE of the MWCD↓50 estimator at $N_p(0, I), t_3$ and t_8. We clearly see that the MWCD↓50 is comparable or better than the MCD50 at t-distributions. For the ARE of the MWCD shape estimators we obtained similar conclusions as for the MWCD location estimators.

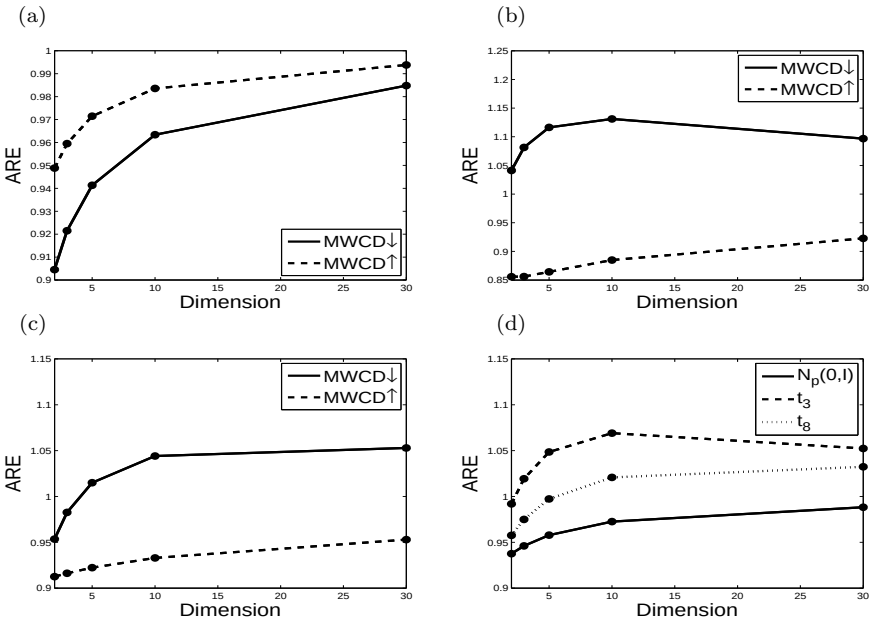

Fig. 1 ARE of the MWCD25 location estimators at (a) normal distribution $N_p(0, I)$; (b) t_3-distribution; and (c) t_8-distribution. (d) ARE of the MWCD↓50 at $N_p(0, I)$, t_3 and t_8

Finally, we consider the ARE of the diagonal elements of the MWCD scatter estimators with $\alpha = 0.25$ which are shown in Figure 2 for respectively $N_p(0, I)$, t_3 and t_8. Note that the ARE of the off-diagonal elements of the scatter is the same as for the shape matrix. From Figure 2a we see that the MWCD↑25 is the most efficient at the normal distribution and comparable to the MCD25 estimator. For t_3 we have the same conclusions as before. For t_8 the MWCD↓25 estimator has the highest ARE ($p > 2$).

Tables 2 and 3 show asymptotic efficiencies for the location, shape and scatter of the MWCD and MCD estimators relative to the sample location, shape and scatter estimators. As is well-known, the efficiencies of the MCD at the normal distribution are low and are directly related to the breakdown point in the sense that a higher breakdown point results in a loss of efficiency. These properties also hold for the MWCD

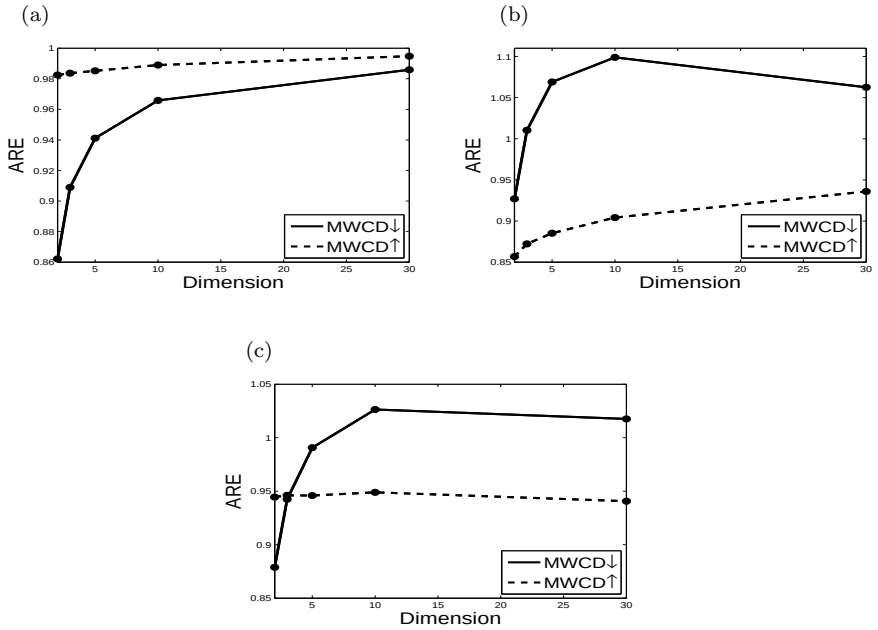

Fig. 2 ARE of the diagonal elements of the MWCD25 scatter estimators at (a) normal distribution $N_p(0, I)$; (b) t_3-distribution; and (c) t_8-distribution

estimators as can be seen from Tables 2 and 3. Note that the efficiencies for the elements of the scatter and shape matrices at the normal distribution would be even lower than those reported in Table 3 for the t_5-distribution. There also exist robust estimators, e.g. MM-estimators (Tatsuoka and Tyler 2000) and τ-estimators (Lopuhaä 1991) that control breakdown point and efficiency at the same time and hence do not suffer from this problem. However, such estimators are typically less appropriate for outlier detection.

We performed a simulation study to investigate the finite-sample performance of the MWCD estimators. The results of this study (not shown) showed that the finite-sample relative efficiencies agree with the asymptotic relative efficiencies.

Table 2 Asymptotic relative efficiencies for the location estimators at the normal distribution $N_p(0, I)$ and t_3-distribution

	$N_p(0, I)$		t_3	
α	0.50	0.25	0.50	0.25
MWCD↓	0.193	0.429	1.117	1.822
MWCD↑	0.200	0.447	1.034	1.442
MCD	0.203	0.466	1.096	1.685

Table 3 Asymptotic relative efficiencies for the diagonal elements of the scatter estimators and off-diagonal elements of the shape estimators at the t_5-distribution

	diagonal		off-diagonal	
α	0.50	0.25	0.50	0.25
MWCD$\downarrow$	0.509	1.311	0.298	0.918
MWCD$\uparrow$	0.513	1.251	0.290	0.836
MCD	0.524	1.358	0.298	0.899

6 Simulations

6.1 Finite-sample robustness

To study the finite-sample robustness of the MWCD estimators, we performed simulations with contaminated data sets. In each simulation we generated 1000 data sets of $N_p(0, I)$ with $p = 3$ and sample sizes $n = 50$, 100, 300 and 500. We considered two typical choices for k, namely $k = \lfloor (n + p + 2)/2 \rfloor$ (corresponding to $\alpha = 0.50$) and $k \approx 0.75n$ (corresponding to $\alpha = 0.25$). To generate contaminated data sets we started with the normally distributed data and then replaced 20% or 40% of the data points x_i by observations with components generated according to $N(s\sqrt{\chi^2_{p,0.99}}, 1.5)$ with $s = 5, 3, 1$. For each simulation we computed the mean squared error and bias of the vectors $\hat{\mu}^{(l)}_{MWCD}$, given by

$$\mathrm{MSE}(\hat{\mu}_{MWCD}) = n \operatorname*{ave}_{1\le j\le p} \operatorname*{ave}_{l} [\{(\hat{\mu}_{MWCD})^{(l)}_j\}^2]$$

$$\mathrm{bias}(\hat{\mu}_{MWCD}) = \sqrt{\operatorname*{ave}_{1\le j\le p} [\{\operatorname*{ave}_{l}(\hat{\mu}_{MWCD})^{(l)}_j\}^2]}.$$

The MSE and bias of diagonal and off-diagonal elements of the shape and scatter matrix were calculated in a similar way.

Tables 4, 5 and 6 show the MSE and the bias for 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 1.5)$ of the estimators MWCD$\downarrow$50, MWCD$\uparrow$50 and MCD50. Table 4 shows that for larger sample sizes the bias of all estimators is close to zero. For $n = 50$ and 100, the MWCD$\downarrow$ location estimator yields the best results. For the estimates of the shape, we see from Table 5 that for small data sets it is better to use the MWCD$\downarrow$ estimator. For the diagonal elements of the scatter estimate (Table 6) the MWCD$\downarrow$ estimator overall shows the best behavior, although all estimates have been considerably affected by the outliers (resulting in large bias and MSE).

Tables 7, 8 and 9 show the MSE and the bias for the MWCD$\downarrow$ and MCD estimator for 20% outliers distributed according to $N(\sqrt{\chi^2_{3,0.99}}, 0.1)$. For $n = 50$ and 100 we see that the bias of the MWCD$\downarrow$ location estimator is smaller than the bias of the MCD location estimator. For $n = 500$ this is no longer true. For the off-diagonal elements of the shape we see the same results. For the diagonal elements of the scatter the MWCD$\downarrow$ has the smallest bias for each n.

Although the MWCD estimators do not improve the efficiency of the MCD for normal data, depending on the type of contamination they can give lower bias.

Table 4 Location estimator: 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 1.5)$

n	50		100		300		500	
	MSE	bias	MSE	bias	MSE	bias	MSE	bias
MWCD↓50	5.777	0.066	2.667	0.005	2.857	0.002	2.981	0.0010
MWCD↑50	8.720	0.109	3.802	0.014	2.688	0.003	2.758	0.0010
MCD50	7.249	0.088	2.844	0.007	2.621	0.003	2.752	0.0014

Table 5 Off-diagonal elements of the shape matrix: 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 1.5)$

n	50		100		300		500	
	MSE	bias	MSE	bias	MSE	bias	MSE	bias
MWCD↓50	33.521	0.184	4.828	0.005	4.943	0.002	4.848	0.004
MWCD↑50	43.787	0.261	10.164	0.028	4.880	0.006	4.744	0.002
MCD50	38.692	0.224	6.576	0.016	4.842	0.005	4.729	0.003

Table 6 Diagonal elements of the scatter matrix: 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 1.5)$

n	50		100		300		500	
	MSE	bias	MSE	bias	MSE	bias	MSE	bias
MWCD↓50	185.043	1.227	92.245	0.832	194.007	0.755	291.939	0.734
MWCD↑50	299.255	1.659	148.520	1.019	236.958	0.841	358.110	0.817
MCD50	227.103	1.400	109.344	0.900	212.955	0.795	327.176	0.780

Table 7 Location estimator: 20% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 0.1)$

n	50		100		300		500	
	MSE	bias	MSE	bias	MSE	bias	MSE	bias
MWCD↓50	7.840	0.075	4.860	0.010	3.836	0.0051	3.870	0.0018
MCD50	7.817	0.082	5.053	0.016	3.565	0.0059	3.666	0.0012
MWCD↓25	2.521	0.020	1.914	0.006	1.522	0.0026	1.722	0.0017
MCD25	3.975	0.049	3.206	0.018	1.419	0.0023	1.584	0.0012

Table 8 Off-diagonal elements of the shape estimator: 20% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 0.1)$

n	50		100		300		500	
	MSE	bias	MSE	bias	MSE	bias	MSE	bias
MWCD↓50	109.075	0.331	39.593	0.044	10.715	0.0071	10.470	0.0024
MCD50	102.312	0.338	39.659	0.053	10.215	0.0075	10.009	0.0037
MWCD↓25	12.687	0.066	5.119	0.010	2.244	0.0036	2.256	0.0036
MCD25	24.864	0.149	16.952	0.051	2.169	0.0026	2.153	0.0033

To illustrate this further we plot in Figure 3 the bias at the different contamination situations for the location estimators versus s where $N(s\sqrt{\chi^2_{3,0.99}}, \theta)$ is the distribu-

Table 9 Diagonal elements of the scatter estimator: 20% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 0.1)$

n	50		100		300		500	
	MSE	bias	MSE	bias	MSE	bias	MSE	bias
MWCD↓50	179.150	0.851	79.430	0.419	47.360	0.288	60.192	0.273
MCD50	175.446	0.904	81.854	0.449	49.427	0.302	62.661	0.284
MWCD↓25	24.412	0.358	20.239	0.329	40.867	0.337	64.735	0.341
MCD25	46.137	0.548	42.995	0.442	50.529	0.381	79.022	0.380

tion of the contaminated components. Figure 3a shows that for 20% less concentrated outliers ($\theta = 1.5$) the MWCD↓ estimator has the lowest bias and the MWCD↑ the highest bias. The decreasing weight function gives a lower weight to points further away, in this case the well spread outliers, so this results in a smaller bias. This bias is also small compared to the MCD which does not make a distinction in weight between the observations in the subset. In case of highly concentrated outliers ($\theta = 0.1$) not far away, the situation is reversed, as seen in Figure 3b. The MWCD↓ estimator now gives the outliers (that is, the most concentrated points) the highest weights and the good observations are given a low weight which results in a high bias. The increasing weight function has the opposite effect and results in a lower bias. If the outliers are further away we get the same relations as with the less concentrated points. For 40% outliers the same conclusions can be made and the effect is even more clear-cut.

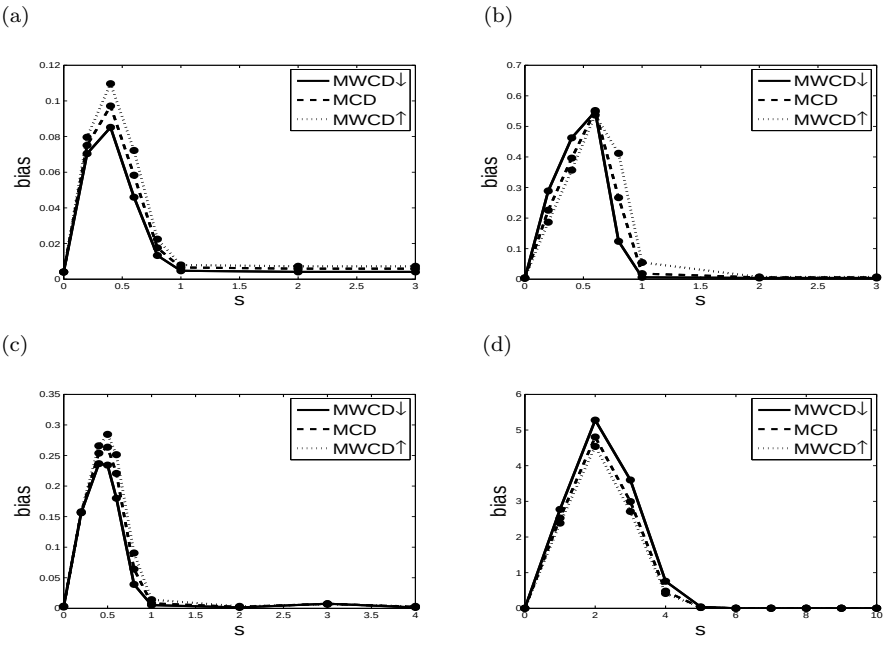

Fig. 3 Bias for the location estimators at ϵ% outliers from $N(s\sqrt{\chi^2_{3,0.99}}, \theta)$, (a) $(\epsilon, \theta, \alpha) = (20\%, 1.5, 0.25)$ (b) $(\epsilon, \theta, \alpha) = (20\%, 0.1, 0.25)$ (c) $(\epsilon, \theta, \alpha) = (40\%, 1.5, 0.50)$ (d) $(\epsilon, \theta, \alpha) = (40\%, 0.1, 0.50)$

7 Examples

7.1 Generated data

To illustrate the use of the two different weight functions in detecting intermediate outliers, we generated a data set with $n = 500$ observations from a multivariate normal distribution $N_p(0, I)$ with $p = 8$. We then replaced 50 of the data points by observations with components generated according to $N(2\sqrt{\chi^2_{8,0.99}}, 0.1)$ and 50 according to $N(0.5\sqrt{\chi^2_{8,0.99}}, 0.1)$. Hence we have a group of strong outliers and a group of intermediate outliers in our data. Figure 4 shows the robust distances of the MWCD↑25 estimator versus those of the MWCD↓25 estimator. The lines correspond to the usual cutoff value $\sqrt{\chi^2_{8,0.975}} = 4.1874$. Both estimators clearly detect the strong outliers lying far away from the majority of the data. However, there are several points that lie above the cutoff for the MWCD↓25 estimator but under the cutoff for the MWCD↑25 estimator. These points correspond exactly to the intermediate outliers in the data. The treatment of the intermediate outliers clearly differs between both estimators. The MWCD↓25 reveals these points because the weight decreases with the distance from the center. On the other hand the MWCD↑25 gives a high weight to these intermediate outliers so that they become masked. Note that a comparable plot is found if we plot the robust distance of the MWCD↑25 versus those of the MCD25 estimator (not shown).

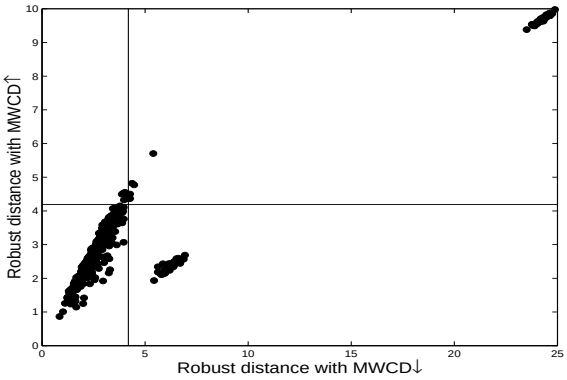

Fig. 4 Generated data: Plot of the robust distances of the MWCD↑25 estimator versus those of the MWCD↓25 estimator

7.2 Ionospheric data

This data set from the Johns Hopkins University Ionosphere database was taken from the "Data Repository" of Hettich and Bay (1999) and has 351 radar measurements on 34 continuous characteristics: real and imaginary parts of the complex responses corresponding to each of 17 pulse numbers. We only look at $n = 225$ 'good' radar

returns showing some type of structure in the ionosphere. As in Maronna and Zamar (2002) variables 1, 2 and 27 are omitted so we are left with $p = 31$ variables. Figure 5 shows a plot of the robust distances of the MWCD↑25 estimator versus the robust distances of the MWCD↓25 estimator. The lines correspond again with the cutoff value $\sqrt{\chi^2_{31,0.975}} = 6.9449$. In Figure 5, 87 observations lie in the upper right corner, meaning that they are detected by both methods as outliers. The critical difference between the weight functions is again in the way that intermediate points are considered. 9 observations lie in the lower right rectangle, which means that they are identified as outliers by MWCD↓25, but not anymore by MWCD↑25. These 9 observations can be considered intermediate outliers, which can be motivated by looking at the sequence of coordinates as done by Maronna and Zamar (2002). They plotted the sequence of coordinates and were able to find 4 characteristic forms to describe the data. These 4 forms can be seen in Figure 6. The outliers detected by both estimators have a much noisier form as shown in Figure 7. We also display in Figure 8 a few typical forms for the intermediate points. We notice that these observations deviate from the pure specimens but are not as aberrant as the outliers. This can explain why these points differ between both estimators. Hence, comparing the outliers that are detected by a decreasing weight function with the outliers detected by an increasing weight function, allows us to identify intermediate outliers.

When comparing both MWCD estimators to MCD25, MWCD↑25 identifies 87 outliers, MCD25 yields 90 outliers (not shown) and MWCD↓25 detects 96 outliers. Hence, the result of the MCD25 estimator, which uses a weight function that is constant on its support, lies in between the two MWCD results, so the MCD25 detects some of the intermediate outliers but not all of them.

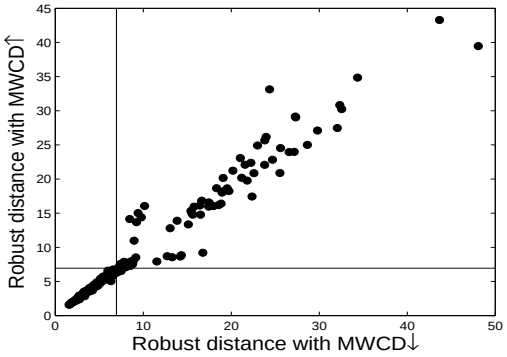

Fig. 5 Ionospheric data: Plot of the robust distances of the MWCD↑ estimator versus those of the MWCD↓ estimator for $\alpha = 0.25$

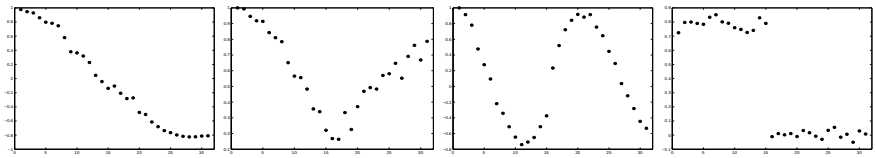

Fig. 6 Ionospheric data: "Pure specimens". Observation 4, observation 32, observation 58 and observation 79 (from left to right).

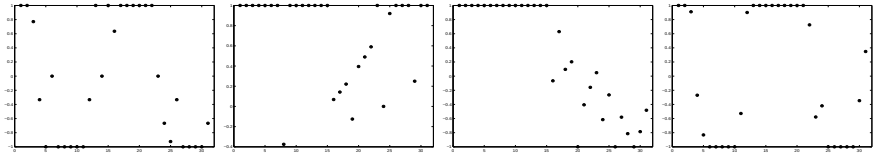

Fig. 7 Ionospheric data: "Outliers". Observation 95, observation 96, observation 41 and observation 27 (from left to right).

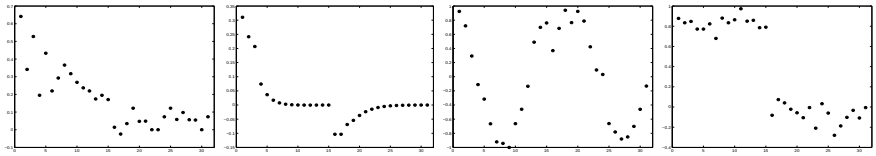

Fig. 8 Ionospheric data: "Intermediate outliers". Observation 67, observation 73, observation 186 and observation 168 (from left to right).

8 Conclusion

We developed a generalization of MCD using weights based on the ranks of the Mahalanobis distances. Similarly to MCD, we used a C-step procedure to construct a fast algorithm to calculate an approximate solution of the MWCD estimators. We showed that the MWCD estimators have the same breakdown point as MCD. We derived influence functions and gave expressions for the asymptotic variances. Comparing the efficiency at several elliptical distributions makes clear that at t-distributions MWCD can give an improvement over MCD, but the efficiencies remain quite low. We also compared the finite-sample robustness in different types of contaminated data sets. For small sample sizes, weighing the observations results in a smaller MSE and bias. For larger sample sizes the situation is less straightforward, but depending on the type of contamination the MWCD has a better bias. Some examples illustrate that using the different weight functions offers possibilities to identify intermediate outliers in the data.

Appendix

The derivations in this appendix are mainly based on the proofs of Van Aelst and Willems (2005) and Agulló et al. (2008).

Proof of Proposition 1 Let $(\hat{\mu}_n, \hat{V}_n) = (\hat{\mu}(X_n), \hat{V}(X_n))$ minimize $\frac{1}{n}\sum_{i=1}^n a_n(R_i)(x_i - m)^T C^{-1}(x_i - m)$ with $\det C = 1$. Let $M = D_n(\hat{\mu}_n, \hat{V}_n)$. We have then

$$\tilde{c} = \frac{1}{n}\sum_{i=1}^n a_n(\tilde{R}_i)(x_i - \hat{\mu}_n)^T \left(\frac{M}{\tilde{c}}\hat{V}_n\right)^{-1}(x_i - \hat{\mu}_n).$$

Then, suppose $(\tilde{\mu}_n, \tilde{\Sigma}_n)$ such that $D_n(\tilde{\mu}_n, \tilde{\Sigma}_n) = \tilde{c}$ and $\det \tilde{\Sigma}_n$ is minimal. This implies $\det \tilde{\Sigma}_n < \det\left(\frac{M}{\tilde{c}}\hat{V}_n\right)$ or $\det\left(\frac{\tilde{c}}{M}\tilde{\Sigma}_n\right) < 1$. Hence there exists a constant $0 < c < 1$ such that $\det(\frac{1}{c}\frac{\tilde{c}}{M}\tilde{\Sigma}_n) = 1$. This implies

$$\frac{1}{n}\sum_{i=1}^n a_n(\tilde{R}_i)(x_i - \tilde{\mu}_n)^T \left(\frac{1}{c}\frac{\tilde{c}}{M}\tilde{\Sigma}_n\right)^{-1}(x_i - \tilde{\mu}_n) = cM < M$$

with $\tilde{R}$ the rank vector of $d^2(\tilde{\mu}_n, \tilde{\Sigma}_n)$ but this contradicts the fact that M corresponds to the minimum of D_n.

Let $(\tilde{\mu}_n, \tilde{\Sigma}_n)$ minimize $\det C$ such that $\frac{1}{n}\sum_{i=1}^n a_n(\tilde{R}_i)(x_i - \tilde{\mu}_n)^T \tilde{\Sigma}_n^{-1}(x_i - \tilde{\mu}_n) = \tilde{c}$. Put $\tilde{V}_n = (\det \tilde{\Sigma}_n)^{-1/p}\tilde{\Sigma}_n$. Then suppose $\hat{V}_n$ such that $\det \hat{V}_n = 1$ and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n a_n(\hat{R}_i)(x_i - \hat{\mu}_n)^T \hat{V}_n^{-1}(x_i - \hat{\mu}_n) \\
&< \frac{1}{n}\sum_{i=1}^n a_n(\tilde{R}_i)(x_i - \tilde{\mu}_n)^T((\det \tilde{\Sigma}_n)^{-1/p}\tilde{\Sigma}_n)^{-1}(x_i - \tilde{\mu}_n) = \tilde{c}(\det \tilde{\Sigma}_n)^{1/p}
\end{aligned}$$

with $\hat{R}$ rank vector of $d^2(\hat{\mu}_n, \hat{V}_n)$ and $\tilde{R}$ rank vector of $d^2(\tilde{\mu}_n, \tilde{V}_n)$. This implies

$$\frac{1}{n}\sum_{i=1}^n a_n(\hat{R}_i)(x_i - \hat{\mu}_n)^T((\det \tilde{\Sigma}_n)^{1/p}\hat{V}_n)^{-1}(x_i - \hat{\mu}_n) < \tilde{c}.$$

Hence there exists a constant $0 < c < 1$ such that

$$\frac{1}{n}\sum_{i=1}^n a_n(\hat{R}_i)(x_i - \hat{\mu}_n)^T((\det \tilde{\Sigma}_n)^{1/p}c\hat{V}_n)^{-1}(x_i - \hat{\mu}_n) = \tilde{c}.$$

But $\det((\det \tilde{\Sigma}_n)^{1/p}c\hat{V}_n) < \det \tilde{\Sigma}_n$ which contradicts that $\tilde{\Sigma}_n$ has minimal determinant. □

Proof of Proposition 2. We first give the following equations similar to those in Agulló et al. (2008) which will be used in the proof. Using properties of traces yields (with $R \in \mathcal{R}$ and $N = \sum_{j=1}^n a_n(R_j)/c_{h^+}$)

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^n a_n(R_j)d_j^2(\hat{\mu}(R), \hat{\Sigma}(R)) &= \frac{1}{N}\sum_{j=1}^n a_n(R_j)(x_j - \hat{\mu}(R))^T \hat{\Sigma}^{-1}(R)(x_j - \hat{\mu}(R)) \\
&= \operatorname{trace}\left(\hat{\Sigma}^{-1}(R)\hat{\Sigma}(R)\right) = p. \qquad (5)
\end{aligned}$$

We also have that

$$\sum_{j=1}^n a_n(R_j)d_j^2(\hat{\mu}(R), \hat{\Sigma}(R)) = (\det \hat{\Sigma}(R))^{-1/p}\sum_{j=1}^n a_n(R_j)d_j^2(\hat{\mu}(R), \hat{V}(R)). \qquad (6)$$

Combining (6) with (5) results in

$$\sum_{j=1}^{n} a_n(R_j) d_j^2(\hat{\mu}(R), \hat{V}(R)) = Np(\det \hat{\Sigma}(R))^{1/p}. \tag{7}$$

We have

$$Q_2 = \sum_{i=1}^{n} a_n(R_{2i}) d_2^2(i) \leq \sum_{i=1}^{n} a_n(R_{1i}) d_2^2(i)$$

because R_{2i} is the rank vector based on $d_2^2(i)$ and a_n is a non-increasing function. Hence it gives the highest weights to the smallest distances, which will result in a smaller sum than any other combination of the weights and the distances would result in. Furthermore,

$$\sum_{i=1}^{n} a_n(R_{1i}) d_2^2(i) \leq \sum_{i=1}^{n} a_n(R_{1i}) d_1^2(i) = Q_1$$

because $\hat{\mu}_2$ and $\hat{V}_2$ minimize $\sum_{i=1}^{n} a_n(R_{1i}) d_i^2(m, C)$. Indeed, suppose that there exist some $m \in \mathbb{R}^p$ and $C \in \mathrm{PDS}(p)$ with $\det C = 1$ such that

$$\sum_{i=1}^{n} a_n(R_{1i}) d_i^2(m, C) < \sum_{i=1}^{n} a_n(R_{1i}) d_i^2(\hat{\mu}_2, \hat{V}_2).$$

Using (7) this implies that

$$\frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i}) d_i^2(m, (\det \hat{\Sigma}_2)^{1/p} C) < p.$$

Hence, there exists a constant $0 < c < 1$ such that $\frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i}) d_i^2(m, c(\det \hat{\Sigma}_2)^{1/p} C) = p$. This can be rewritten as

$$\mathrm{trace}\left(c^{-1} (\det \hat{\Sigma}_2)^{-1/p} C^{-1} \frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i})(x-m)(x-m)^T \right) = p.$$

If we then use the following maximum determinant result: if one maximizes $\det A$ over all positive semi-definite symmetric matrices of size p with $\mathrm{trace}(A) = p$, then the solution is $A = I$. We then have

$$\det\left(\frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i})(x-m)(x-m)^T \right) \leq \det(c(\det \hat{\Sigma}_2)^{1/p} C) = c^p \det \hat{\Sigma}_2.$$

Hence,

$$\det\left(\frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i})(x-\hat{\mu}_2)(x-\hat{\mu}_2)^T \right) = \det \hat{\Sigma}_2$$
$$\leq \det\left(\frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i})(x-m)(x-m)^T \right) \leq c^p \det \hat{\Sigma}_2$$

which is a contradiction so we have that

$$\sum_{i=1}^{n} a_n(R_{1i}) d_i^2(m, C) \geq \sum_{i=1}^{n} a_n(R_{1i}) d_i^2(\hat{\mu}_2, \hat{V}_2).$$

We have equality if and only if

$$\sum_{i=1}^{n} a_n(R_{2i}) d_2^2(i) = \sum_{i=1}^{n} a_n(R_{1i}) d_2^2(i) = \sum_{i=1}^{n} a_n(R_{1i}) d_1^2(i).$$

Using (7) on the last equality this implies that

$$\frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i}) d_i^2(\hat{\mu}_1, (\det \hat{\Sigma}_2)^{1/p} \hat{V}_1) = p.$$

This can be rewritten as

$$\text{trace} \left((\det \hat{\Sigma}_2)^{-1/p} \hat{V}_1^{-1} \frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i})(x - \hat{\mu}_1)(x - \hat{\mu}_1)^T \right) = p.$$

Using the maximum determinant result we get

$$\det \hat{\Sigma}_2 \leq \det \left(\frac{1}{N} \sum_{i=1}^{n} a_n(R_{1i})(x - \hat{\mu}_1)(x - \hat{\mu}_1)^T \right) \leq \det((\det \hat{\Sigma}_2)^{1/p} \hat{V}_1) = \det \hat{\Sigma}_2.$$

This implies that $\hat{\mu}_2 = \hat{\mu}_1$. Using $\text{trace} \left((\det \hat{\Sigma}_2)^{-1/p} \hat{V}_1^{-1} \hat{\Sigma}_2 \right) = p$ and $\det \left((\det \hat{\Sigma}_2)^{-1/p} \hat{V}_1^{-1} \hat{\Sigma}_2 \right) = 1$ the maximum determinant result implies $\hat{V}_2 = \hat{V}_1$. □

Proof of Proposition 3 We first prove that

$$\varepsilon_n^*(\hat{\mu}_{MWCD}, X_n) \geq \frac{\min(n - k + 1, k - k(X_n))}{n}.$$

We will show that there exist $\bar{M}$ and α which only depend on X_n, such that for every X_n' obtained by replacing at most $s = \min(n-k+1, k-k(X_n))-1$ observations from X_n we have that $\|\hat{\mu}_{MWCD}(X_n')\| \leq \bar{M}$, $\lambda_1(\hat{\Sigma}_{MWCD}(X_n')) \leq \alpha$ and $\lambda_p(\hat{\Sigma}_{MWCD}(X_n')) > 0$. The norm we use here is the L_2 norm. Let us denote

$$\mathcal{E}(t, C) = \{x : (x - t)^T C^{-1} (x - t) \leq p/a_m\}$$

with $t \in \mathbb{R}^p$, $C \in PDS(p)$ and $a_m = \min_{a_n(R_i) > 0} a_n(R_i)$. Consider the ellipsoids $\mathcal{E}(0, cI)$ such that

$$\sum_{i=1}^{n} a_n(R_i) x_i^T I c^{-1} x_i = p$$

with R_i the rank vector of $\|x_i\|^2$. Choose a ranking $R^* \in \mathcal{R}$ such that the largest distances $x_i^T x_i$ get the highest weights and denote $c_{\max}$ the corresponding constant such that $\sum_{i=1}^{n} a_n(R_i^*) x_i^T I c_{\max}^{-1} x_i = p$.

With R_i' the rank vector of $\|x_i'\|^2$, it holds that

$$\sum_{i=1}^{n} a_n(R_i')(x_i')^T x_i' \leq \sum_{i=1}^{n} a_n(R_i^*) x_i^T x_i$$

because $n - s \geq k$ implies that X_n' still contains k data points of the original X_n. For the ellipsoid $\mathcal{E}(0, c_m I)$ such that

$$\sum_{i=1}^{n} a_n(R_i')(x_i')^T I c_m^{-1} x_i' = p$$

it then holds that $\det(c_m I) \leq \det(c_{\max} I)$. It follows that for the optimal solution $\mathcal{E}(\hat{\mu}_{MWCD}(X'_n), \hat{\Sigma}_{MWCD}(X'_n))$ that satisfies

$$\sum_{i=1}^{n} a_n(\tilde{R}_i)(x'_i - \hat{\mu}_{MWCD}(X'_n))^T \hat{\Sigma}_{MWCD}(X'_n)^{-1}(x'_i - \hat{\mu}_{MWCD}(X'_n)) = p \qquad (8)$$

with $\tilde{R}$ the rank vector corresponding to the distances $d_i(\hat{\mu}_{MWCD}(X'_n), \hat{\Sigma}_{MWCD}(X'_n))$ we must have that $\det(\hat{\Sigma}_{MWCD}(X'_n)) \leq c_{\max}^p = V$. Condition (8) implies that the ellipsoid $\mathcal{E}(\hat{\mu}_{MWCD}(X'_n), \hat{\Sigma}_{MWCD}(X'_n))$ contains a subcollection of at least k points of X'_n. Because $s \leq k - k(X_n) - 1$ this subcollection contains at least $k(X_n) + 1$ data points of the original X_n in general position. Using lemma 3.1 of Lopuhaä and Rousseeuw (1991) if $||\hat{\mu}_{MWCD}(X'_n)|| > \bar{M}$ we obtain that $\det(\hat{\Sigma}_{MWCD}(X'_n)) > V$, yielding a contradiction. We have thus shown that $\|\hat{\mu}_{MWCD}(X'_n)\| \leq \bar{M}$. Moreover, since $\mathcal{E}(\hat{\mu}_{MWCD}(X'_n), \hat{\Sigma}_{MWCD}(X'_n))$ contains at least $k(X_n) + 1$ original data points in general position we know that there exists a constant $\beta > 0$ such that $\lambda_j(\hat{\Sigma}_{MWCD}(X'_n)) \geq \beta$ for all $j = 1, \ldots, p$ and $\det(\hat{\Sigma}_{MWCD}(X'_n)) < V$ then implies that there exists a constant $0 < \alpha < \infty$ (depending on β and V) such that $\lambda_1(\hat{\Sigma}_{MWCD}(X'_n)) \leq \alpha$.

Let us now prove that $\varepsilon_n^*(\hat{\mu}_{MWCD}, X_n), \varepsilon_n^*(\hat{\Sigma}_{MWCD}, X_n) \leq \min(n - k + 1, k - k(X_n))/n$. Suppose we contaminate $s = \min(n - k + 1, k - k(X_n))$ points of X_n to obtain X'_n. Suppose first that $s = n - k + 1$. Let $\mathcal{E}(t, c)$ be an ellipsoid that satisfies

$$\sum_{i=1}^{n} a_n(R_i)(x'_i - t)^T C^{-1}(x'_i - t) = p$$

with R_i rank of $(x'_i - t)^T C^{-1}(x'_i - t)$ for $i = 1, \ldots, n$. Because $n - s = k - 1$ there exists at least one contaminated point that belongs to $\mathcal{E}(t, C)$. By letting $\|x\| \to \infty$ for the replaced points we can make sure that at least one of the eigenvalues of C goes to infinity. (If $\mathcal{E}(t, C)$ contains only replaced points, then letting $\|x\| \to \infty$ in different directions assures that $\det(C) \to \infty$). Therefore both $\hat{\mu}_{MWCD}(X'_n)$ and $\hat{\Sigma}_{MWCD}(X'_n)$ break down in this case.

Suppose $s = k - k(X_n)$. Denote $\tilde{J} \subset \{1, \ldots, n\}$ the set of indices corresponding to the $k(X_n)$ observations from X_n lying on a hyperplane of $\mathbb{R}^p$. Then there exist an $\alpha \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}$ such that $\alpha^T x_j - \gamma = 0$ for all $j \in \tilde{J}$. There exists a $m \in \mathbb{R}^p$ such that $m^T \alpha = \gamma$ which implies $\alpha^T (x_j - m) = 0$ for $j \in \tilde{J}$. Therefore for $j \in \tilde{J}$ we have that $x_j - m \in S$ where S is a $(p-1)$dimensional subspace of $\mathbb{R}^p$. Now take a $d \in \mathbb{R}^p$ with $\|d\| = 1$ such that $d \in S$. Now replace $s = k - k(X_n)$ observations of X_n, not lying on S by $(m + \lambda d)$ for some arbitrarily chosen $\lambda \in \mathbb{R}$. Denote J_0 the set of indices corresponding to the outliers. It follows that for the s outliers $x_j - m - \lambda d = 0$ and for the $k(X_n)$ points on S we have that $x_j - m - \lambda d \in S$. Denote $\{e_1, \ldots, e_{p-1}\}$ an orthonormal basis of S and e_p a normed vector orthogonal to S. Denote $P = [e_1 \ldots e_p]$. Consider $C = P \Lambda P^T$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. For the $k(X_n)$ points we have that $x_j - m - \lambda d \in S$, thus there exist for each $j \in \tilde{J}$ coefficients $\zeta_1, \ldots, \zeta_{p-1}$ such that

$x_j - m - \lambda d = \sum_{i=1}^{p-1} \zeta_i e_i$. Therefore

$$\begin{aligned}
&(x_j - (m+\lambda d))^T C^{-1}(x_j - (m+\lambda d)) \\
&= \sum_{i=1}^{p-1} \zeta_i e_i^T \left(\sum_{j=1}^{p} \lambda_j^{-1} e_j e_j^T \right) \sum_{i=1}^{p-1} \zeta_i e_i \\
&= \left(\sum_{i=1}^{p-1} \zeta_i \lambda_i^{-1} e_i^T \right) \sum_{i=1}^{p-1} \zeta_i e_i = \sum_{i=1}^{p-1} \zeta_i^2 \lambda_i^{-1}.
\end{aligned}$$

Now $\sum_{i=1}^{n} a_n(R_i)(x_i - (m+\lambda d))^T C^{-1}(x_i - (m+\lambda d)) =$

$$\sum_{s \text{ outliers}} + \sum_{k(X_n) \text{ on } S} + \sum_{\text{remainder}}$$

with R_i the rank of $(x_i - (m+\lambda d))^T C^{-1}(x_i - (m+\lambda d))$ for $i = 1, \ldots, n$. Fix λ and choose $(\lambda_1, \ldots, \lambda_{p-1})$ appropriately such that $\sum_{k(X_n) \text{ on } S} \to p$. For the remainder we can write

$$\sum_{\text{remainder}} a_n(R_i)(x_i - (m+\lambda d))^T C^{-1}(x_i - (m+\lambda d)) = \sum_{\text{remainder}} a_n(R_i) \left(\sum_{i=1}^{p} \zeta_i^2 / \lambda_i \right).$$

If we let $\lambda_p \to 0$ then the corresponding distances $\to \infty$ and will surely be the largest distances getting weight 0. The constructed solution has $\det(C) = \lambda_1 \cdots \lambda_p \to 0$. By letting $\lambda \to \infty$ we thus obtain that both $\hat{\mu}_{MWCD}(X_n')$ and $\hat{\Sigma}_{MWCD}(X_n')$ break down. □

Proof of Proposition 4 First, we consider a non-increasing weight function h^+. For any (m, C) denote $V_{m,C} = (\det \Sigma_{m,C})^{-1/p} \Sigma_{m,C}$ such that $\det V_{m,C} = 1$. Using properties of traces yields with $N = \int h^+(G(d_x^2(m,C)))dH(x)/c_{h^+}$

$$\begin{aligned}
&\frac{1}{N} E_H[h^+(G(d_x^2(m,C)))d_x^2(\mu_{m,C}, \Sigma_{m,C})] \\
&= \frac{1}{N} E_H[h^+(G(d_x^2(m,C)))(x - \mu_{m,C})^T \Sigma_{m,C}^{-1}(x - \mu_{m,C})] \\
&= \frac{1}{N} E_H[\text{trace}\left(h^+(G(d_x^2(m,C)))(x - \mu_{m,C})^T \Sigma_{m,C}^{-1}(x - \hat{\mu}_{m,C}) \right)] \\
&= \frac{1}{N} E_H[\text{trace}\left(h^+(G(d_x^2(m,C)))(x - \mu_{m,C})(x - \mu_{m,C})^T \Sigma_{m,C}^{-1} \right)] \\
&= \frac{1}{N} \text{trace}\left(E_H[h^+(G(d_x^2(m,C)))(x - \mu_{m,C})(x - \mu_{m,C})^T \Sigma_{m,C}^{-1}] \right) \\
&= \text{trace}(\Sigma_{m,C} \Sigma_{m,C}^{-1}) = p. \qquad (9)
\end{aligned}$$

We also have that

$$\begin{aligned}
&E_H[h^+(G(d_x^2(m,C)))d_x^2(\mu_{m,C}, \Sigma_{m,C})] \\
&= (\det \Sigma_{m,C})^{-1/p} E_H[h^+(G(d_x^2(m,C)))d_x^2(\mu_{m,C}, V_{m,C})]. \qquad (10)
\end{aligned}$$

Combining (10) with (9) results in

$$E_H[h^+(G(d_x^2(m,C)))d_x^2(\mu_{m,C}, V_{m,C})] = Np(\det \Sigma_{m,C})^{1/p}. \qquad (11)$$

We prove that for any $(\tilde{\mu}, \tilde{V}) \in \underset{m,C;\det C=1}{\operatorname{argmin}} D(m, C)$ that $\tilde{\mu} = \mu_{\tilde{\mu},\tilde{\Sigma}}$ and $\tilde{\Sigma} = \Sigma_{\tilde{\mu},\tilde{\Sigma}}$ and $(\tilde{\mu}, \tilde{\Sigma})$ minimizes $\underset{m=\mu_{m,C};C=\Sigma_{m,C}}{\operatorname{argmin}} \det C$. Because Proposition 1 also holds in the functional form, $(\tilde{\mu}, \tilde{\Sigma})$ minimizes $\det \Sigma$ such that $\frac{1}{N} D(\mu, \Sigma) = p$. This implies

$$\frac{1}{N} E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{\Sigma})))(x - \tilde{\mu})^T \tilde{\Sigma}^{-1}(x - \tilde{\mu})] = p.$$

This can be rewritten as

$$\operatorname{trace}\left(\tilde{\Sigma}^{-1} \frac{1}{N} E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{\Sigma})))(x - \tilde{\mu})(x - \tilde{\mu})^T]\right) = p.$$

Using the maximum determinant result the determinant of this matrix is maximized by 1. Or this can be written as

$$\det\left(\frac{1}{N} E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{\Sigma})))(x - \tilde{\mu})(x - \tilde{\mu})^T]\right) \leq \det \tilde{\Sigma}.$$

It holds that

$$\begin{aligned}
\det \Sigma_{\tilde{\mu},\tilde{\Sigma}} &= \det\left(\frac{1}{N} E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{\Sigma})))(x - \mu_{\tilde{\mu},\tilde{\Sigma}})(x - \mu_{\tilde{\mu},\tilde{\Sigma}})^T]\right) \\
&\leq \det\left(\frac{1}{N} E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{\Sigma})))(x - \tilde{\mu})(x - \tilde{\mu})^T]\right) \qquad (12) \\
&\leq \det \tilde{\Sigma}. \qquad (13)
\end{aligned}$$

Because h^+ is a non-increasing function it holds

$$\begin{aligned}
&\frac{1}{N} E_H[h^+(G(d_x^2(\mu_{\tilde{\mu},\tilde{\Sigma}}, \Sigma_{\tilde{\mu},\tilde{\Sigma}})))(x - \mu_{\tilde{\mu},\tilde{\Sigma}})^T \Sigma_{\tilde{\mu},\tilde{\Sigma}}^{-1}(x - \mu_{\tilde{\mu},\tilde{\Sigma}})] \\
\leq\ &\frac{1}{N} E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{\Sigma})))(x - \mu_{\tilde{\mu},\tilde{\Sigma}})^T \Sigma_{\tilde{\mu},\tilde{\Sigma}}^{-1}(x - \mu_{\tilde{\mu},\tilde{\Sigma}})] = p.
\end{aligned}$$

Thus there exists a constant $0 < c \leq 1$ such that

$$\frac{1}{N} E_H[h^+(G(d_x^2(\mu_{\tilde{\mu},\tilde{\Sigma}}, \Sigma_{\tilde{\mu},\tilde{\Sigma}})))(x - \mu_{\tilde{\mu},\tilde{\Sigma}})^T (c\Sigma_{\tilde{\mu},\tilde{\Sigma}})^{-1}(x - \mu_{\tilde{\mu},\tilde{\Sigma}})] = p.$$

This implies $\det \tilde{\Sigma} \leq \det c\Sigma_{\tilde{\mu},\tilde{\Sigma}} \leq \det \Sigma_{\tilde{\mu},\tilde{\Sigma}}$. Hence together with (13) this results in $\det \tilde{\Sigma} = \det \Sigma_{\tilde{\mu},\tilde{\Sigma}}$ and because (12) becomes an equality also $\tilde{\mu} = \mu_{\tilde{\mu},\tilde{\Sigma}}$. Because

$$\operatorname{trace}\left(\tilde{\Sigma}^{-1} \frac{1}{N} E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{\Sigma})))(x - \mu_{\tilde{\mu},\tilde{\Sigma}})(x - \mu_{\tilde{\mu},\tilde{\Sigma}})^T]\right) = \operatorname{trace}(\tilde{\Sigma}^{-1}\Sigma_{\tilde{\mu},\tilde{\Sigma}}) = p$$

and $\det(\tilde{\Sigma}^{-1}\Sigma_{\tilde{\mu},\tilde{\Sigma}}) = 1$ it follows from the maximum determinant result that $\tilde{\Sigma} = \Sigma_{\tilde{\mu},\tilde{\Sigma}}$.

For any $m = \mu_{m,C}$ and $C = \Sigma_{m,C}$ with $V = V_{m,C} = (\det C)^{-1/p} C$

$$\begin{aligned}
&E_H[h^+(G(d_x^2(\tilde{\mu}, \tilde{V})))d_x^2(\tilde{\mu}, \tilde{V})] \\
\leq\ &E_H[h^+(G(d_x^2(m, C)))d_x^2(m, V)] = E_H[h^+(G(d_x^2(m, C)))d_x^2(\mu_{m,C}, V_{m,C})].
\end{aligned}$$

Using (11) the inequality can be rewritten as

$$Np(\det \Sigma_{\tilde{\mu},\tilde{V}})^{1/p} \leq Np(\det \Sigma_{m,C})^{1/p}$$

so we obtain $\det \Sigma_{\tilde{\mu},\tilde{V}} \leq \det \Sigma_{m,C}$ for all $m = \mu_{m,C}$ and $C = \Sigma_{m,C}$ hence we conclude that $(\tilde{\mu}, \tilde{\Sigma}) \in \underset{m=\mu_{m,C};C=\Sigma_{m,C}}{\operatorname{argmin}} \det C$.

For an increasing weight function we may argue as follows. From differentiating the objective function $D(m, C)$, we easily obtain that the minimizing m and C can be represented as a weighted mean and covariance, on condition that $D(m, C)$ is continuously differentiable in this point. The latter condition will not always be satisfied. However, if the distribution H is continuous, then it holds that $D(m, C)$ is continuously differentiable. Hence, we must have that the solution of minimizing $D(m, C)$ does satisfy the weighted mean and covariance representation. □

Proof of Proposition 5 First of all, due to equivariance, we may assume that $\mu = 0$ and $\Sigma = I_p$. We will denote $F = F_{0,I_p}$. Hence, we are left to show that for any MWCD solution we have that $\mu_{MWCD}(F) = 0$ and $\Sigma_{MWCD}(F) = I_p$. μ_{MWCD} is the weighted mean based solely on the ellipse $\mathcal{E} = \{x \in \mathbb{R}^p; (x - \mu_{MWCD})^T \Sigma_{MWCD}^{-1}(x - \mu_{MWCD}) \leq q_\alpha\}$ implying that

$$\int_{\mathcal{E}} w(d^2(x))(x - \mu_{MWCD})^T dF(x) = 0 \tag{14}$$

with $d^2(x) = (x - \mu_{MWCD})^T \Sigma_{MWCD}^{-1}(x - \mu_{MWCD})$ and $w = h^+ \circ \tilde{G}$. Suppose that $\mu_{MWCD} \neq 0$. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of Σ_{MWCD} and $v_1, \ldots, v_p$ the corresponding eigenvectors. There will be at least one $1 \leq j \leq p$ such that $\mu_{MWCD}^T v_j \neq 0$. Fix this j. From (14) it follows that we should have

$$\int_{\mathcal{E}} v_j^T \mu_{MWCD} w(d^2(x))(x - \mu_{MWCD})^T v_j dF(x) = 0. \tag{15}$$

Set $d = (d_1, \ldots, d_p)^T := \mu_{MWCD}$. Since x is spherically symmetrically distributed we may assume w.l.o.g. that $\Sigma_{MWCD} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ as well as $v_j = (1, 0, \ldots, 0)$. For every $d_1 - \sqrt{c\lambda_1} \leq x_1 \leq d_1 + \sqrt{c\lambda_1}$ denote

$$\mathcal{E}(x_1) = \left\{ (x_2, \ldots, x_p) \in \mathbb{R}^{p-1} | \sum_{j=2}^{p} \frac{(x_j - d_j)^2}{\lambda_j} \leq c - \frac{(x_1 - d_1)^2}{\lambda_1} \right\}$$

where $c := q_\alpha > 0$. Then we have

$$\begin{aligned}
I &= \int_{\mathcal{E}} w(d^2(x))(x - \mu_{MWCD})^T v_j dF(x) \\
&= \int_{d_1 - \sqrt{c\lambda_1}}^{d_1 + \sqrt{c\lambda_1}} \int_{\mathcal{E}(x_1)} w\left(\sum_{j=1}^{p} \frac{(x_j - d_j)^2}{\lambda_j} \right) (x_1 - d_1) g(x_1^2 + \ldots + x_p^2) dx_1 \ldots dx_p \\
&= \int_{-\sqrt{c\lambda_1}}^{\sqrt{c\lambda_1}} t \int_{\mathcal{E}(d_1 + t)} w\left(\frac{t^2}{\lambda_1} + \sum_{j=2}^{p} \frac{(x_j - d_j)^2}{\lambda_j} \right) \times \\
&\quad g\left((d_1 + t)^2 + x_2^2 + \cdots + x_p^2 \right) dx_2 \ldots dx_p dt.
\end{aligned}$$

Since $\mathcal{E}(d_1 + t) = \mathcal{E}(d_1 - t)$ and $w(d^2(d_1 + t, x_2, \ldots, x_p)) = w(d^2(d_1 - t, x_2, \ldots, x_p))$ it follows that

$$\begin{aligned}
I = \int_0^{\sqrt{c\lambda_1}} t \int_{\mathcal{E}(d_1 + t)} w\left(\tfrac{t^2}{\lambda_1} + \textstyle\sum_{j=2}^{p} \tfrac{(x_j - d_j)^2}{\lambda_j} \right) \Big[g\left((d_1 + t)^2 + x_2^2 + \cdots + x_p^2 \right) \\
- g\left((d_1 - t)^2 + x_2^2 + \cdots + x_p^2 \right) \Big] dx_2 \ldots dx_p dt.
\end{aligned}$$

If $d_1 > 0$ we have $(d_1 + t)^2 + x_2^2 + \cdots + x_p^2 > (d_1 - t)^2 + x_2^2 + \cdots + x_p^2$ and since g is strictly decreasing this implies $I < 0$. Similarly, we can show that $d_1 < 0$ yields $I > 0$. Therefore, we have shown that $v_j^T \mu_{MWCD} > 0$ implies $I < 0$ and if $v_j^T \mu_{MWCD} < 0$ then $I > 0$. Hence, we obtain $\int_{\mathcal{E}} v_j^T \mu_{MWCD} w(d^2(x))(x - \mu_{MWCD})^T v_j dF_{\mu,\Sigma}(x) < 0$ which contradicts (15) so we conclude that $\mu_{MWCD} = 0$.

We now show the Fisher consistency of Σ_{MWCD}. The derivation is similar to the proof of Lemma 3 in Butler et al. (1993). We have already shown that $\mathcal{E} = \{x \in \mathbb{R}^p; x^T \Sigma_{MWCD}^{-1} x \leq q_\alpha\}$. As before, we may assume that Σ_{MWCD} is a diagonal matrix Λ with diagonal elements $\lambda_1, \ldots, \lambda_p$. We have

$$\Lambda = \lambda \int_{\mathcal{E}} w(d^2(x)) x x^T g(x_1^2 + \cdots + x_p^2) dx,$$

for some $\lambda > 0$. On writing $y = \Lambda^{-1/2} x$, it is sufficient to show that all solutions of

$$I_p = \lambda' \int_{\mathcal{E}} w(\|y\|^2) y y^T g\left(\sum_{i=1}^p \lambda_i y_i^2\right) dy$$

for some $\lambda' > 0$ satisfy $\lambda_1 = \ldots = \lambda_p$.

We have

$$\int_{\mathcal{E}} w(\|y\|^2) y_1^2 g\left(\sum_{i=1}^p \lambda_i y_i^2\right) dy = \int_{\mathcal{E}} w(\|y\|^2) y_2^2 g\left(\sum_{i=1}^p \lambda_i y_i^2\right) dy \tag{16}$$

and hence

$$\int_{\mathcal{E}} w(\|y\|^2)(y_1^2 - y_2^2)\left[g\left(\lambda_1 y_1^2 + \lambda_2 y_2^2 + \sum_{i=3}^p \lambda_i y_i^2\right) - g\left(\lambda_2 y_1^2 + \lambda_1 y_2^2 + \sum_{i=3}^p \lambda_i y_i^2\right)\right] dy = 0 \tag{17}$$

as may be seen by interchanging the roles of y_1 and y_2. Suppose $\lambda_1 > \lambda_2$. Then if $y_1^2 > y_2^2$ it follows that $\lambda_1 y_1^2 + \lambda_2 y_2^2 > \lambda_2 y_1^2 + \lambda_1 y_2^2$. Similarly, if $y_1^2 < y_2^2$ then $\lambda_1 y_1^2 + \lambda_2 y_2^2 < \lambda_2 y_1^2 + \lambda_1 y_2^2$. Thus if $\lambda_1 > \lambda_2$ the integral in (17) is always non-positive and strictly negative at some y_1, y_2. This contradicts (16) showing $\lambda_1 = \lambda_2$ and in general $\lambda_1 = \ldots = \lambda_p$.

Finally the consistency factor c_{h+} then makes sure that $\Sigma_{MWCD}(F) = I_p$. □

Proof of Proposition 6. Consider the contaminated distribution $F_\varepsilon = (1-\varepsilon)F_0 + \varepsilon \Delta_{x_0}$ and denote $\mu_\varepsilon := \mu_{MWCD}(F_\varepsilon)$ and $\Sigma_\varepsilon := \Sigma_{MWCD}(F_\varepsilon)$. We have then that

$$\mu_\varepsilon = \frac{\int w(d_{F_\varepsilon}^2(x)) x dF_\varepsilon(x)}{\int w(d_{F_\varepsilon}^2(x)) dF_\varepsilon(x)}$$

is an MWCD solution. Differentiating w.r.t. ε and evaluating at 0 yields

$$\begin{aligned} IF(x_0; \mu_{MWCD}, F_0) = &\left(\int w(d_{F_0}^2(x)) dF_0(x)\right)^{-1} \frac{\partial}{\partial \varepsilon} \int w(d_{F_\varepsilon}^2(x)) x dF_\varepsilon(x)_{|\varepsilon=0} \\ &+ \frac{\partial}{\partial \varepsilon}\left[\left(\int w(d_{F_\varepsilon}^2(x)) dF_\varepsilon(x)\right)^{-1}\right]_{|\varepsilon=0} \int w(d_{F_0}^2(x)) x dF_0(x). \end{aligned}$$

By symmetry of F_0

$$\int w(d^2_{F_0}(x))xdF_0(x) = 0$$

and

$$\int w(d^2_{F_0}(x))dF_0(x) = \int w(||x||^2)g(x^Tx)dx = c_1$$

or $c_1 = \frac{2\pi^{p/2}}{\Gamma(p/2)} \int w(r^2)g(r^2)r^{p-1}dr$. Hence, we obtain

$$\begin{aligned}
&IF(x_0; \mu_{MWCD}, F_0) \\
&= \frac{1}{c_1}\frac{\partial}{\partial\varepsilon}\int w(d^2_{F_\varepsilon}(x))xdF_\varepsilon(x)_{|\varepsilon=0} \\
&= \frac{1}{c_1}\frac{\partial}{\partial\varepsilon}\left((1-\varepsilon)\int w(d^2_{F_\varepsilon}(x))xdF_0(x) + \varepsilon w(d^2_{F_0}(x_0))x_0\right)_{|\varepsilon=0} \\
&= \frac{1}{c_1}\Big(w(x_0^Tx_0)x_0 - \int w(d^2_{F_\varepsilon}(x))xdF_0(x)_{|\varepsilon=0} \\
&\quad + (1-\varepsilon)\frac{\partial}{\partial\varepsilon}\int w(d^2_{F_\varepsilon}(x))xdF_0(x)_{|\varepsilon=0}\Big) \\
&= \frac{1}{c_1}\Big(w(x_0^Tx_0)x_0 + \frac{\partial}{\partial\varepsilon}\int w(d^2_{F_\varepsilon}(x))xdF_0(x)_{|\varepsilon=0}\Big). \qquad (18)
\end{aligned}$$

We now simplify the last term:

$$\int w(d^2_{F_\varepsilon}(x))xdF_0(x) = \int w(d^2_{F_\varepsilon}(x))xg(x^Tx)dx.$$

Using the transformation $v = \Sigma_\varepsilon^{-1/2}(x-\mu_\varepsilon)$ yields

$$\begin{aligned}
\mathcal{I}_1(\varepsilon) &:= \int w(d^2_{F_\varepsilon}(x))xg(x^Tx)dx \\
&= \det(\Sigma_\varepsilon)^{1/2}\int w(v^Tv)(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)g((\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)^T(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon))dv.
\end{aligned}$$

If we rewrite this expression in polar coordinates $v = re(\theta)$ then $r \in [0, \sqrt{q_\alpha(\varepsilon)}]$. This is because the function w differs only from zero when $d^2_\varepsilon(x) = (x-\mu_\varepsilon)^T\Sigma_\varepsilon^{-1}(x-\mu_\varepsilon) \leq q_\alpha(\varepsilon)$ where $q_\alpha(\varepsilon) = (D^2_{F_\varepsilon})^{-1}(1-\alpha)$ with $D^2_{F_\varepsilon}(t) = P_{F_\varepsilon}(d^2_\varepsilon(x) \leq t)$. $e(\theta) \in S^{p-1}$ and $\theta = (\theta_1, \ldots, \theta_{p-1}) \in \Theta = [0,\pi[\times\ldots\times[0,\pi[\times[0,2\pi[$, yields

$$\begin{aligned}
\mathcal{I}_1(\varepsilon) = \det(\Sigma_\varepsilon)^{1/2}\int_0^{\sqrt{q_\alpha(\varepsilon)}}\int_\Theta J(\theta,r)w(re(\theta)^Tre(\theta))(r\Sigma_\varepsilon^{1/2}e(\theta)+\mu_\varepsilon)\times \\
g((r\Sigma_\varepsilon^{1/2}e(\theta)+\mu_\varepsilon)^T(r\Sigma_\varepsilon^{1/2}e(\theta)+\mu_\varepsilon))drd\theta,
\end{aligned}$$

where $J(r,\theta)$ is the Jacobian of the transformation into polar coordinates. Using Leibniz' formula to this expression and the symmetry of F_0 results in

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}\mathcal{I}_1(\varepsilon)_{|\varepsilon=0} \\
&= \int_{||v||^2\leq q_\alpha}\frac{\partial}{\partial\varepsilon}\left(w(v^Tv)(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)g((\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)^T(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon))\right)_{|\varepsilon=0}dv
\end{aligned}$$

because

$$\frac{1}{2}\text{trace}(IF(x_0;\Sigma_{MWCD},F_0))\int_{||v||^2\leq q_\alpha} w(v^Tv)vg(v^Tv)dv = 0$$

and

$$\frac{\partial(\sqrt{q_\alpha(\varepsilon)})}{\partial\varepsilon}_{|_{\varepsilon=0}}\int_\Theta J(\theta,\sqrt{q_\alpha})w(q_\alpha)\sqrt{q_\alpha}Ie(\theta)g(q_\alpha)d\theta = 0.$$

We obtain for the derivative on the right hand side

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}\{w(v^Tv)(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)g((\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)^T(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon))\}_{|_{\varepsilon=0}}\\
&= w(v^Tv)IF(x_0;\Sigma_{MWCD}^{1/2},F_0)vg(v^Tv) + w(v^Tv)IF(x_0;\mu_{MWCD},F_0)^Tg(v^Tv)\\
&+ 2w(v^Tv)vg'(v^Tv)\{v^TIF(x_0,\Sigma_{MWCD}^{1/2},F_0)v + v^TIF(x_0,\mu_{MWCD},F_0)\}.
\end{aligned}$$

Since $\int_{||v||^2\leq q_\alpha} w(v^Tv)vg(v^Tv)dv$ and
$\int_{||v||^2\leq q_\alpha} w(v^Tv)vg'(v^Tv)v^TIF(x_0;\Sigma_{MWCD}^{1/2},F_0)vdv$ are zero due to symmetry of F_0, we see that the terms including $IF(x_0;\Sigma_{MWCD}^{1/2},F_0)$ give a zero contribution to the integral. Therefore,

$$\begin{aligned}
\frac{\partial}{\partial\varepsilon}\mathcal{I}_1(\varepsilon)_{|_{\varepsilon=0}} &= IF(x_0;\mu_{MWCD},F_0)\int_{||v||^2\leq q_\alpha} w(v^Tv)g(v^Tv)dv\\
&\quad + 2\int_{||v||^2\leq q_\alpha} w(v^Tv)g'(v^Tv)vv^Tdv IF(x_0;\mu_{MWCD},F_0)\\
&= [c_1+2c_2]IF(x_0;\mu_{MWCD},F_0)
\end{aligned}$$

where $c_2 = \int_{||v||^2\leq q_\alpha} w(v^Tv)g'(v^Tv)v_1^2dv$ can be rewritten by using polar coordinates (see end). We now have that

$$\frac{\partial}{\partial\varepsilon}\int_{d_\varepsilon^2(x)\leq q_\alpha(\varepsilon)} w(d_{F_\varepsilon}^2(x))xdF_0(x) = [c_1+2c_2]IF(x_0;\mu_{MWCD},F_0). \qquad (19)$$

Substituting (19) in (18) yields

$$c_1IF(x_0;\mu_{MWCD},F_0) = w(x_0^Tx_0)x_0I(||x_0||^2\leq q_\alpha) + [c_1+2c_2]IF(x_0;\mu_{MWCD},F_0)$$

which gives the final result

$$IF(x;\mu_{MWCD},F_0) = \frac{1}{-2c_2}w(\|x\|^2)xI(||x||^2\leq q_\alpha)$$

$$\begin{aligned}
c_2 &= \int_{||v||^2\leq q_\alpha} w(v^Tv)g'(v^Tv)v_1^2dv\\
&= \frac{1}{p}\int_{||v||^2\leq q_\alpha} w(v^Tv)g'(v^Tv)(v^Tv)dv\\
&= \frac{1}{p}\int_0^{\sqrt{q_\alpha}}\frac{2\pi^{p/2}}{\Gamma(p/2)}w(r^2)g'(r^2)r^{p-1}r^2dr\\
&= \frac{\pi^{p/2}}{\Gamma(p/2+1)}\int_0^{\sqrt{q_\alpha}} w(r^2)g'(r^2)r^{p+1}dr.
\end{aligned}$$

Similarly, the influence function of the scatter matrix part can be derived (see also Croux and Haesbroeck 1999). We have that

$$\Sigma_\varepsilon = c_{h^+} \frac{\int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T dF_\varepsilon(x)}{\int w(d^2_{F_\varepsilon}(x)) dF_\varepsilon(x)}$$

is an MWCD solution. Differentiating with respect to ε and evaluating at 0 yields

$$\begin{aligned}
&IF(x_0; \Sigma_{MWCD}, F_0) \\
&= \left(\int w(d^2_{F_0}(x)) dF_0(x) \right)^{-1} \frac{\partial}{\partial \varepsilon} c_{h^+} \int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T dF_\varepsilon(x)_{|_{\varepsilon=0}} \\
&+ \frac{\partial}{\partial \varepsilon} \left[\left(\int w(d^2_{F_\varepsilon}(x)) dF_\varepsilon(x) \right)^{-1} \right]_{|_{\varepsilon=0}} c_{h^+} \int w(d^2_{F_0}(x)) x x^T dF_0(x). \qquad (20)
\end{aligned}$$

The second term in (20) is zero, so only the first term remains, for which we use:

$$\int w(d^2_{F_0}(x)) dF_0(x) = c_1.$$

Therefore, we get:

$$\begin{aligned}
&IF(x_0; \Sigma_{MWCD}, F_0) \\
&= \frac{c_{h^+}}{c_1} \frac{\partial}{\partial \varepsilon} \int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T dF_\varepsilon(x)_{|_{\varepsilon=0}} \\
&= \frac{c_{h^+}}{c_1} \frac{\partial}{\partial \varepsilon} \left[\int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T (1-\varepsilon) dF_0(x) + \varepsilon w(d^2_{F_0}(x_0)) x_0 x_0^T \right]_{|_{\varepsilon=0}} \\
&= \frac{c_{h^+}}{c_1} \frac{\partial}{\partial \varepsilon} \left[\int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T dF_0(x) \right. \\
&\quad \left. - \varepsilon \int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T dF_0(x) + \varepsilon w(d^2_{F_0}(x_0)) x_0 x_0^T \right]_{|_{\varepsilon=0}} \\
&= \frac{c_{h^+}}{c_1} \left(\frac{\partial}{\partial \varepsilon} \int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T dF_0(x)_{|_{\varepsilon=0}} \right. \\
&\quad \left. - \int w(d^2_{F_\varepsilon}(x))(x-\mu_\varepsilon)(x-\mu_\varepsilon)^T dF_0(x)_{|_{\varepsilon=0}} + w(d^2_{F_0}(x_0)) x_0 x_0^T \right) \\
&= \frac{c_{h^+}}{c_1} \left(\frac{\partial}{\partial \varepsilon} \int w(d^2_{F_\varepsilon}(x)) x x^T dF_0(x)_{|_{\varepsilon=0}} - \int w(d^2_{F_0}(x)) x x^T dF_0(x) + w(d^2_{F_0}(x_0)) x_0 x_0^T \right) \\
&= \frac{1}{c_3} \left(\frac{\partial}{\partial \varepsilon} \int w(d^2_{F_\varepsilon}(x)) x x^T dF_0(x)_{|_{\varepsilon=0}} - \int w(d^2_{F_0}(x)) x x^T dF_0(x) + w(d^2_{F_0}(x_0)) x_0 x_0^T \right).
\end{aligned}$$

We have that

$$\begin{aligned}
-\frac{1}{c_3}\int w(d^2_{F_0}(x))xx^T dF_0(x) &= -\frac{1}{c_3}\int w(\|x\|^2)xx^T g(x^Tx)dx \\
&= \left(-\frac{1}{c_3}\frac{1}{p}\int w(\|x\|^2)x^Txg(x^Tx)dx\right) I \\
&= \left(-\frac{1}{c_3}\frac{1}{p}\frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty w(r^2)r^2r^{p-1}g(r^2)dr\right) I \\
&= \left(-\frac{1}{c_3}\frac{1}{p}\frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty w(r^2)r^{p+1}g(r^2)dr\right) I \\
&= -I.
\end{aligned}$$

Hence

$$IF(x_0;\Sigma_{MWCD},F_0) = \frac{1}{c_3}\frac{\partial}{\partial\varepsilon}\int w(d^2_{F_\varepsilon}(x))xx^T dF_0(x)_{|_{\varepsilon=0}} - I + \frac{w(\|x_0\|^2)x_0x_0^T}{c_3}.$$

We rewrite this as

$$\int w(d^2_{F_\varepsilon}(x))xx^T dF_0(x) = \int w(d^2_{F_\varepsilon}(x))xx^T g(x^Tx)dx.$$

Using again the transformation $v = \Sigma_\varepsilon^{-1/2}(x-\mu_\varepsilon)$ we obtain that

$$\begin{aligned}
\mathcal{I}_2(\varepsilon) &:= \int w(d^2_{F_\varepsilon}(x))xx^T g(x^Tx)dx \\
&= \det(\Sigma_\varepsilon)^{1/2}\int w(v^Tv)(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)^T \times \\
&\quad g((\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)^T(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon))dv.
\end{aligned}$$

As before we rewrite this expression in polar coordinates $v = re(\theta)$ with $r\in[0,\sqrt{q_\alpha(\varepsilon)}]$, $e(\theta)\in S^{p-1}$ and $\theta=(\theta_1,\ldots,\theta_{p-1})\in\Theta=[0,\pi[\times\cdots\times[0,\pi[\times[0,2\pi[$ which yields

$$\begin{aligned}
\mathcal{I}_2(\varepsilon) = \det(\Sigma_\varepsilon)^{1/2}\int_0^{\sqrt{q_\alpha(\varepsilon)}}\int_\Theta \Big[J(\theta,r)w(re(\theta)^Tre(\theta))(r\Sigma_\varepsilon^{1/2}e(\theta)+\mu_\varepsilon)\times \\
(r\Sigma_\varepsilon^{1/2}e(\theta)+\mu_\varepsilon)^Tg((r\Sigma_\varepsilon^{1/2}e(\theta)+\mu_\varepsilon)^T(r\Sigma_\varepsilon^{1/2}e(\theta)+\mu_\varepsilon))\Big]drd\theta. \qquad (21)
\end{aligned}$$

Applying Leibniz formula to (21) and using the symmetry of F_0 results in:

$$\begin{aligned}
\frac{\partial\mathcal{I}_2(\varepsilon)}{\partial\varepsilon}_{|_{\varepsilon=0}} &= \frac{1}{2}\text{trace}(IF(x_0;\Sigma_{MWCD},F_0))H(q_\alpha)I \\
&\quad + \frac{\partial\sqrt{q_\alpha(\varepsilon)}}{\partial\varepsilon}_{|_{\varepsilon=0}} q_\alpha w(q_\alpha)g(q_\alpha)d_1I \\
&\quad + \int_{\|v\|^2\le q_\alpha}\frac{\partial}{\partial\varepsilon}\Big(w(v^Tv)(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)^T\times \\
&\quad g((\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon)^T(\Sigma_\varepsilon^{1/2}v+\mu_\varepsilon))\Big)_{|_{\varepsilon=0}}dv \qquad (22)
\end{aligned}$$

with $d_1 = \int_\Theta J(\theta, \sqrt{q_\alpha}) e_1^2(\theta) d\theta = \frac{1}{p} \int_\Theta J(\theta, \sqrt{q_\alpha}) d\theta$ and $H(q_\alpha) = \int_{\|v\|^2 \leq q_\alpha} w(\|v\|^2) v v^T g(v^T v) dv = c_3$. The last term of (22) can be worked out as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \varepsilon} \left(w(v^T v)(\Sigma_\varepsilon^{1/2} v + \mu_\varepsilon)(\Sigma_\varepsilon^{1/2} v + \mu_\varepsilon)^T g((\Sigma_\varepsilon^{1/2} v + \mu_\varepsilon)^T (\Sigma_\varepsilon^{1/2} v + \mu_\varepsilon)) \right)_{|\varepsilon=0} dv \\
&= \frac{1}{2} w(v^T v) \Big\{ IF(x_0; \Sigma_{MWCD}, F_0) v v^T + v v^T IF(x_0; \Sigma_{MWCD}, F_0) \\
&\quad + 2 IF(x_0; \mu_{MWCD}, F_0) v^T + 2 v IF(x_0; \mu_{MWCD}, F_0)^T \Big\} g(v^T v) \\
&+ w(v^T v) v v^T g'(v^T v) \Big\{ v^T IF(x_0; \Sigma_{MWCD}, F_0) v + 2 v^T IF(x_0; \mu_{MWCD}, F_0) \Big\}. \quad (23)
\end{aligned}$$

Note that since $\int_{\|v\|^2 \leq q_\alpha} v w(v^T v) g(v^T v) dv$ and $\int_{\|v\|^2 \leq q_\alpha} v v^T w(v^T v) g'(v^T v) v dv$ are zero, the terms in (23) including $IF(x_0, \mu_{MWCD}, F_0)$ give a zero contribution to the integral in (22). We still need to compute the term $\frac{\partial \sqrt{q_\alpha(\varepsilon)}}{\partial \varepsilon}_{|\varepsilon=0}$ of (22). Using

$$c_1 = \int w(d^2_{F_\varepsilon}(x)) dF_\varepsilon(x) = (1-\varepsilon) \int w(d^2_{F_\varepsilon}(x)) dF_0(x) + \varepsilon w(d^2_{F_\varepsilon}(x_0)) I(d^2_\varepsilon(x_0) \leq q_\alpha(\varepsilon))$$

and differentiating both sides with respect to ε yields

$$\begin{aligned}
0 &= \frac{\partial}{\partial \varepsilon} \int w(d^2_{F_\varepsilon}(x)) dF_0(x)_{|\varepsilon=0} - \int w(d^2_{F_0}(x)) dF_0(x) + w(\|x_0\|^2) I(\|x_0\|^2 \leq q_\alpha) \\
&= \frac{1}{2} \text{trace}(IF(x_0; \Sigma_{MWCD}, F_0)) c_1 + \frac{\partial \sqrt{q_\alpha(\varepsilon)}}{\partial \varepsilon}_{|\varepsilon=0} w(q_\alpha) g(q_\alpha) \int_\Theta J(\theta, \sqrt{q_\alpha}) d\theta \\
&+ \int_{\|v\|^2 \leq q_\alpha} w(v^T v) g'(v^T v) v^T IF(x_0; \Sigma_{MWCD}, F_0) v dv - c_1 \\
&+ w(\|x_0\|^2) I(\|x_0\|^2 \leq q_\alpha). \quad (24)
\end{aligned}$$

The third term in (24) equals, using the symmetry of F_0, $c_2 \text{trace}(IF(x_0; \Sigma_{MWCD}, F_0))$. This leads to

$$\frac{\partial \sqrt{q_\alpha(\varepsilon)}}{\partial \varepsilon}_{|\varepsilon=0} = \frac{c_1 - w(\|x_0\|^2) I(\|x_0\|^2 \leq q_\alpha) - \text{trace}(IF(x_0; \Sigma_{MWCD}, F_0))(c_2 + \frac{c_1}{2})}{w(q_\alpha) g(q_\alpha) p d_1}. \quad (25)$$

So inserting (25) and (23) in (22) yields

$$\begin{aligned}
&IF(x_0; \Sigma_{MWCD}, F_0) \\
&= \frac{1}{2} \text{trace}(IF(x_0; \Sigma_{MWCD}, F_0)) I \\
&+ \frac{1}{c_3} \frac{q_\alpha}{p} \left(c_1 - w(\|x_0\|^2) I(\|x_0\|^2 \leq q_\alpha) - \text{trace}(IF(x_0; \Sigma_{MWCD}, F_0))(c_2 + \frac{c_1}{2}) \right) I \\
&+ \frac{1}{2c_3} \int_{\|v\|^2 \leq q_\alpha} w(v^T v) \Big(IF(x_0; \Sigma_{MWCD}, F_0) v v^T \\
&\quad + v v^T IF(x_0; \Sigma_{MWCD}, F_0) \Big) g(v^T v) dv \\
&+ \frac{1}{c_3} \int_{\|v\|^2 \leq q_\alpha} w(v^T v) v v^T g'(v^T v) v^T IF(x_0; \Sigma_{MWCD}, F_0) v dv \\
&- I + \frac{1}{c_3} w(\|x_0\|^2) x_0 x_0^T. \quad (26)
\end{aligned}$$

In order to give elementwise expressions for the influence function we use results from Croux and Haesbroeck (1999) to see:

$$\begin{aligned}&\frac{1}{2}\sum_{k=1}^{p}\left\{IF(x_0;\Sigma_{ik},F_0)\int_{\|v\|^2\le q_\alpha} w(v^Tv)v_kv_jg(v^Tv)dv\right.\\&\left.+IF(x_0;\Sigma_{kj},F_0)\int_{\|v\|^2\le q_\alpha} w(v^Tv)v_iv_kg(v^Tv)dv\right\}\\&=H(q_\alpha)IF(x_0;\Sigma_{ij},F_0)\end{aligned}$$

for every $1\le i,j\le p$ and

$$\begin{aligned}&\sum_{k=1}^{p}\sum_{l=1}^{p}IF(x_0;\Sigma_{kl},F_0)\int_{\|v\|^2\le q_\alpha} v_iv_jv_kv_lw(v^Tv)g'(v^Tv)dv\\&=\begin{cases}c_4\text{trace}(IF(x_0;\Sigma_{MWCD},F_0))+(c_5-c_4)IF(x_0;\Sigma_{ii},F_0) & i=j\\ 2c_4IF(x_0;\Sigma_{ij},F_0) & i\ne j\end{cases}\end{aligned}$$

where $c_4=\int_{\|v\|^2\le q_\alpha} v_i^2v_j^2w(v^Tv)g'(v^Tv)dv$ and $c_5=\int_{\|v\|^2\le q_\alpha} v_i^4w(v^Tv)g'(v^Tv)dv$. From (26) the influence function for the off-diagonal elements is straightforwardly obtained,

$$IF(x_0;\Sigma_{ij},F_0)=\frac{1}{c_3}(2c_4+H(q_\alpha))IF(x_0;\Sigma_{ij},F_0)+\frac{1}{c_3}x_{0i}x_{0j}w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha)$$

hence

$$IF(x_0;\Sigma_{ij},F_0)=-\frac{1}{2c_4}x_{0i}x_{0j}w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha).$$

For the diagonal elements we get

$$\begin{aligned}&IF(x_0;\Sigma_{jj},F_0)\\&=\frac{1}{2}\text{trace}(IF(x_0;\Sigma_{MWCD},F_0))\\&+\frac{1}{c_3}\frac{q_\alpha}{p}\left\{c_1-w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha)-\text{trace}(IF(x_0;\Sigma_{MWCD},F_0))(c_2+\frac{c_1}{2})\right\}\\&+\frac{1}{c_3}\{c_5-c_4+H(q_\alpha)\}IF(x_0;\Sigma_{jj},F_0)+\frac{c_4}{c_3}\text{trace}(IF(x_0;\Sigma_{MWCD},F_0))\\&-1+\frac{x_{0j}^2w(\|x_0\|^2)}{c_3}I(\|x_0\|^2\le q_\alpha).\end{aligned}$$

Using

$$b_1=\frac{1}{c_3}(c_4-c_5)\text{ and }b_2=\frac{1}{2}+\frac{1}{c_3}\left[c_4-\frac{q_\alpha}{p}(c_2+\frac{c_1}{2})\right]$$

leads to

$$\begin{aligned}&b_1IF(x_0;\Sigma_{jj},F_0)-b_2\text{trace}(IF(x_0;\Sigma_{MWCD},F_0))\\&=-1+\frac{w(\|x_0\|^2)}{c_3}x_{0j}^2I(\|x_0\|^2\le q_\alpha)+\frac{1}{c_3}\frac{q_\alpha}{p}\left\{c_1-w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha)\right\}.\end{aligned}\tag{27}$$

Taking the sum of the diagonal terms in (27) yields an expression for the trace of the influence function:

$$\begin{aligned}\text{trace}(IF(x_0;\Sigma_{MWCD},F_0)) = (b_1-pb_2)^{-1}\Big\{&\frac{1}{c_3}\|x_0\|^2 w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha)\\ &+p\{\frac{1}{c_3}\frac{q_\alpha}{p}(c_1-w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha))-1\}\Big\}.\end{aligned} \tag{28}$$

Using (28) in (27) yields

$$\begin{aligned}IF(x_0;\Sigma_{jj},F_0) = \frac{1}{b_1}\Big\{&\frac{1}{c_3}x_{0j}^2 w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha)\\ &+\frac{b_2}{b_1-pb_2}\frac{1}{c_3}\|x_0\|^2 w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha)\\ &+\frac{b_1}{b_1-pb_2}\left[\frac{1}{c_3}\frac{q_\alpha}{p}(c_1-w(\|x_0\|^2)I(\|x_0\|^2\le q_\alpha))-1\right]\Big\}.\end{aligned}$$

The function $R(\|x\|)$ is defined as:

$$\begin{aligned}R(\|x\|) = \frac{1}{b_1}\Big\{&\frac{b_2}{b_1-pb_2}\frac{1}{c_3}\|x\|^2 w(\|x\|^2)I(\|x\|^2\le q_\alpha)\\ &+\frac{b_1}{b_1-pb_2}\left[\frac{1}{c_3}\frac{q_\alpha}{p}(c_1-w(\|x\|^2)I(\|x\|^2\le q_\alpha))-1\right]\Big\}.\end{aligned} \tag{29}$$

□

References

1. Agulló J, Croux C, Van Aelst S (2008) The multivariate least trimmed squares estimator. Journal of Multivariate Analysis 99: 311 – 338
2. Butler RW, Davies PL, Jhun M (1993) Asymptotics for the minimum covariance determinant estimator. The Annals of Statistics 21: 1385 – 1400
3. Croux C, Dehon C (2002) Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. Revue de Statistique Appliquée 50: 5 – 26
4. Croux C, Haesbroeck G (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71: 161 – 190
5. Croux C, Haesbroeck G (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika 87: 603 – 618
6. Davies PL (1987) Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. The Annals of Statistics 15: 1269 – 1292
7. Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL (eds.) A festschrift for Erich Lehmann. Wadsworth, Belmont, California, pp. 157 – 184
8. Hadi AS, Luceño A (1997) Maximum trimmed likelihood estimators: a unified approach, examples and algorithms. Computational Statistics & Data Analysis 25: 251 – 272
9. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York
10. Hettich S, Bay SD (1999) "The UCI KDD Archive" [http://kdd.ics.uci.edu], University of California, Irvine, Dept. of Information and Computer Science
11. Hössjer O (1994) Rank-based estimates in the linear model with high breakdown point. Journal of the American Statistical Association 89: 149 – 158
12. Kent JT, Tyler DE (1996) Constrained M-estimation for multivariate location and scatter. The Annals of Statistics 24: 1346 – 1370

13. Lopuhaä HP (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance. The Annals of Statistics 17: 1662 – 1683
14. Lopuhaä HP (1991) Multivariate τ-estimators for location and scatter. The Canadian Journal of Statistics 19: 307 – 321
15. Lopuhaä HP, Rousseeuw PJ (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics 19: 229 – 248
16. Maronna RA (1976) Robust M-estimators of multivariate location and scatter. The Annals of Statistics 4: 51 – 67
17. Maronna RA, Zamar RH (2002) Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44: 307 – 317
18. Masicek L (2004) Optimality of the least weighted squares estimator. Kybernetika 40: 715 – 734
19. Pison G, Rousseeuw PJ, Filzmoser P, Croux C (2003) Robust factor analysis. Journal of Multivariate Analysis 84: 145 – 172
20. Pison G, Van Aelst S (2004) Diagnostic plots for robust multivariate methods. Journal of Computational and Graphical Statistics 13: 310 – 329
21. Rousseeuw PJ (1984) Least median of squares regression. Journal of the American Statistical Association 79: 871 – 880
22. Rousseeuw PJ, Van Aelst S, Van Driessen K, Agulló J (2004) Robust multivariate regression. Technometrics 46: 293 – 305
23. Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41: 212 – 223
24. Salibian-Barrera M, Van Aelst S, Willems G (2006) Principal components analysis based on multivariate MM-estimators with fast and robust bootstrap. Journal of the American Statistical Association 101: 1198 – 1211
25. Taskinen S, Croux C, Kankainen A, Ollila E, Oja H (2006) Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. Journal of Multivariate Analysis 97: 359 – 384
26. Tatsuoka KS, Tyler DE (2000) On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. The Annals of Statistics 28: 1219 – 1243
27. Van Aelst S, Willems G (2005) Multivariate regression S-estimators for robust estimation and inference. Statistica Sinica 15: 981 – 1001
28. Vandev DL, Neykov NM (1998) About regression estimators with high breakdown point. Statistics 32: 111 – 129
29. Visek JA (2001) Regression with high breakdown point. In: ROBUST'2000, Proceedings of the 11th conference on robust statistics, pp. 324 – 356